FaceForensics++: Learning to Detect Manipulated Facial Images
Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, Matthias Niessner; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1-11
The rapid progress in synthetic image generation and manipulation has now come to a point where it raises significant concerns for the implications towards society. At best, this leads to a loss of trust in digital content, but could potentially cause further harm by spreading false information or fake news. This paper examines the realism of state-of-the-art image manipulations, and how difficult it is to detect them, either automatically or by humans. To standardize the evaluation of detection methods, we propose an automated benchmark for facial manipulation detection. In particular, the benchmark is based on Deep-Fakes, Face2Face, FaceSwap and NeuralTextures as prominent representatives for facial manipulations at random compression level and size. The benchmark is publicly available and contains a hidden test set as well as a database of over 1.8 million manipulated images. This dataset is over an order of magnitude larger than comparable, publicly available, forgery datasets. Based on this data, we performed a thorough analysis of data-driven forgery detectors. We show that the use of additional domain-specific knowledge improves forgery detection to unprecedented accuracy, even in the presence of strong compression, and clearly outperforms human observers.
****
DeepVCP: An End-to-End Deep Neural Network for Point Cloud Registration
Weixin Lu, Guowei Wan, Yao Zhou, Xiangyu Fu, Pengfei Yuan, Shiyu Song; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 12-21
We present DeepVCP - a novel end-to-end learning-based 3D point cloud registration framework that achieves comparable registration accuracy to prior state-of-the-art geometric methods. Different from other keypoint based methods where a RANSAC procedure is usually needed, we implement the use of various deep neural network structures to establish an end-to-end trainable network. Our keypoint detector is trained through this end-to-end structure and enables the system to avoid the interference of dynamic objects, leverages the help of sufficiently salient features on stationary objects, and as a result, achieves high robustness. Rather than searching the corresponding points among existing points, the key contribution is that we innovatively generate them based on learned matching probabilities among a group of candidates, which can boost the registration accuracy. We comprehensively validate the effectiveness of our approach using both the KITTI dataset and the Apollo-SouthBay dataset. Results demonstrate that our method achieves comparable registration accuracy and runtime efficiency to the state-of-the-art geometry-based methods, but with higher robustness to inaccurate initial poses. Detailed ablation and visualization analysis are included to further illustrate the behavior and insights of our network. The low registration error and high robustness of our method make it attractive to the substantial applications relying on the point cloud registration task.
****
Shape Reconstruction Using Differentiable Projections and Deep Priors
Matheus Gadelha, Rui Wang, Subhransu Maji; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 22-30
We investigate the problem of reconstructing shapes from noisy and incomplete projections in the presence of viewpoint uncertainities. The problem is cast as an optimization over the shape given measurements obtained by a projection operator and a prior. We present differentiable projection operators for a number of reconstruction problems which when combined with the deep image prior or shape prior allows efficient inference through gradient descent. We apply our method on a variety of reconstruction problems, such as tomographic reconstruction from a few samples, visual hull reconstruction incorporating view uncertainties, and 3D shape reconstruction from noisy depth maps. Experimental results show that our approach is effective for such shape reconstruction problems, without requiring any task-specific training.

****
Fine-Grained Segmentation Networks: Self-Supervised Segmentation for Improved Long-Term Visual Localization
Mans Larsson, Erik Stenborg, Carl Toft, Lars Hammarstrand, Torsten Sattler, Fredrik Kahl; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 31-41
Long-term visual localization is the problem of estimating the camera pose of a given query image in a scene whose appearance changes over time. It is an important problem in practice that is, for example, encountered in autonomous driving. In order to gain robustness to such changes, long-term localization approaches often use segmantic segmentations as an invariant scene representation, as the semantic meaning of each scene part should not be affected by seasonal and other changes. However, these representations are typically not very discriminative due to the very limited number of available classes. In this paper, we propose a novel neural network, the Fine-Grained Segmentation Network (FGSN), that can be used to provide image segmentations with a larger number of labels and can be trained in a self-supervised fashion. In addition, we show how FGSNs can be trained to output consistent labels across seasonal changes. We show through extensive experiments that integrating the fine-grained segmentations produced by our FGSNs into existing localization algorithms leads to substantial improvements in localization performance.
****
SANet: Scene Agnostic Network for Camera Localization
Luwei Yang, Ziqian Bai, Chengzhou Tang, Honghua Li, Yasutaka Furukawa, Ping Tan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 42-51
This paper presents a scene agnostic neural architecture for camera localization, where model parameters and scenes are independent from each other.Despite recent advancement in learning based methods, most approaches require training for each scene one by one, not applicable for online applications such as SLAM and robotic navigation, where a model must be built on-the-fly.Our approach learns to build a hierarchical scene representation and predicts a dense scene coordinate map of a query RGB image on-the-fly given an arbitrary scene. The 6D camera pose of the query image can be estimated with the predicted scene coordinate map. Additionally, the dense prediction can be used for other online robotic and AR applications such as obstacle avoidance. We demonstrate the effectiveness and efficiency of our method on both indoor and outdoor benchmarks, achieving state-of-the-art performance.
****
Total Denoising: Unsupervised Learning of 3D Point Cloud Cleaning
Pedro Hermosilla, Tobias Ritschel, Timo Ropinski; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 52-60
We show that denoising of 3D point clouds can be learned unsupervised, directly from noisy 3D point cloud data only. This is achieved by extending recent ideas from learning of unsupervised image denoisers to unstructured 3D point clouds. Unsupervised image denoisers operate under the assumption that a noisy pixel observation is a random realization of a distribution around a clean pixel value, which allows appropriate learning on this distribution to eventually converge to the correct value. Regrettably, this assumption is not valid for unstructured points: 3D point clouds are subject to total noise, i.e. deviations in all coordinates, with no reliable pixel grid. Thus, an observation can be the realization of an entire manifold of clean 3D points, which makes the quality of a naive extension of unsupervised image denoisers to 3D point clouds unfortunately only little better than mean filtering. To overcome this, and to enable effective and unsupervised 3D point cloud denoising, we introduce a spatial prior term, that steers converges to the unique closest out of the many possible modes on the manifold. Our results demonstrate unsupervised denoising performance similar to that of supervised learning with clean data when given enough training examples - whereby we do not need any pairs of noisy and clean training data.
****

Hierarchical Self-Attention Network for Action Localization in Videos
Rizard Renanda Adhi Pramono, Yie-Tarng Chen, Wen-Hsien Fang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 61-70
This paper presents a novel Hierarchical Self-Attention Network (HISAN) to generate spatial-temporal tubes for action localization in videos. The essence of HISAN is to combine the two-stream convolutional neural network (CNN) with hierarchical bidirectional self-attention mechanism, which comprises of two levels of bidirectional self-attention to efficaciously capture both of the long-term temporal dependency information and spatial context information to render more precise action localization. Also, a sequence rescoring (SR) algorithm is employed to resolve the dilemma of inconsistent detection scores incurred by occlusion or background clutter. Moreover, a new fusion scheme is invoked, which integrates not only the appearance and motion information from the two-stream network, but also the motion saliency to mitigate the effect of camera motion. Simulations reveal that the new approach achieves competitive performance as the state-of-the-art works in terms of action localization and recognition accuracy on the widespread UCF101-24 and J-HMDB datasets.
****

Goal-Driven Sequential Data Abstraction
Umar Riaz Muhammad, Yongxin Yang, Timothy M. Hospedales, Tao Xiang, Yi-Zhe Song; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 71-80
Automatic data abstraction is an important capability for both benchmarking machine intelligence and supporting summarization applications. In the former one asks whether a machine can `understand' enough about the meaning of input data to produce a meaningful but more compact abstraction. In the latter this capability is exploited for saving space or human time by summarizing the essence of input data. In this paper we study a general reinforcement learning based framework for learning to abstract sequential data in a goal-driven way. The ability to define different abstraction goals uniquely allows different aspects of the input data to be preserved according to the ultimate purpose of the abstraction. Our reinforcement learning objective does not require human-defined examples of ideal abstraction. Importantly our model processes the input sequence holistically without being constrained by the original input order. Our framework is also domain agnostic -- we demonstrate applications to sketch, video and text data and achieve promising results in all domains.
****

Jointly Aligning Millions of Images With Deep Penalised Reconstruction Congealing
Roberto Annunziata, Christos Sagonas, Jacques Cali; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 81-90
Extrapolating fine-grained pixel-level correspondences in a fully unsupervised manner from a large set of misaligned images can benefit several computer vision and graphics problems, e.g. co-segmentation, super-resolution, image edit propagation, structure-from-motion, and 3D reconstruction. Several joint image alignment and congealing techniques have been proposed to tackle this problem, but robustness to initialisation, ability to scale to large datasets, and alignment accuracy seem to hamper their wide applicability. To overcome these limitations, we propose an unsupervised joint alignment method leveraging a densely fused spatial transformer network to estimate the warping parameters for each image and a low-capacity auto-encoder whose reconstruction error is used as an auxiliary measure of joint alignment. Experimental results on digits from multiple versions of MNIST (i.e., original, perturbed, affNIST and infiMNIST) and faces from LFW, show that our approach is capable of aligning millions of images with high accuracy and robustness to different levels and types of perturbation. Moreover, qualitative and quantitative results suggest that the proposed method outperforms state-of-the-art approaches both in terms of alignment quality and robustness to initialisation.
****

Drop to Adapt: Learning Discriminative Features for Unsupervised Domain Adaptati

on
Seungmin Lee, Dongwan Kim, Namil Kim, Seong-Gyun Jeong; Proceedings of the IE EE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 91-100
Recent works on domain adaptation exploit adversarial training to obtain domain-invariant feature representations from the joint learning of feature extractor a nd domain discriminator networks. However, domain adversarial methods render sub optimal performances since they attempt to match the distributions among the dom ains without considering the task at hand. We propose Drop to Adapt (DTA), which leverages adversarial dropout to learn strongly discriminative features by enfo rcing the cluster assumption. Accordingly, we design objective functions to supp ort robust domain adaptation. We demonstrate efficacy of the proposed method on various experiments and achieve consistent improvements in both image classifica tion and semantic segmentation tasks. Our source code is available at https://gi thub.com/postBG/DTA.pytorch.
****

NLNL: Negative Learning for Noisy Labels
Youngdong Kim, Junho Yim, Juseung Yun, Junmo Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 101-110
Convolutional Neural Networks (CNNs) provide excellent performance when used for image classification. The classical method of training CNNs is by labeling imag es in a supervised manner as in "input image belongs to this label" (Positive Le arning; PL), which is a fast and accurate method if the labels are assigned corr ectly to all images. However, if inaccurate labels, or noisy labels, exist, trai ning with PL will provide wrong information, thus severely degrading performance . To address this issue, we start with an indirect learning method called Negati ve Learning (NL), in which the CNNs are trained using a complementary label as i n "input image does not belong to this complementary label." Because the chances of selecting a true label as a complementary label are low, NL decreases the ri sk of providing incorrect information. Furthermore, to improve convergence, we e xtend our method by adopting PL selectively, termed as Selective Negative Learni ng and Positive Learning (SelNLPL). PL is used selectively to train upon expecte d-to-be-clean data, whose choices become possible as NL progresses, thus resulti ng in superior performance of filtering out noisy data. With simple semi-supervi sed training technique, our method achieves state-of-the-art accuracy for noisy data classification, proving the superiority of SelNLPL's noisy data filtering a bility.
****

Adversarial Robustness vs. Model Compression, or Both?
Shaokai Ye, Kaidi Xu, Sijia Liu, Hao Cheng, Jan-Henrik Lambrechts, Huan Zha ng, Aojun Zhou, Kaisheng Ma, Yanzhi Wang, Xue Lin; Proceedings of the IEEE/C VF International Conference on Computer Vision (ICCV), 2019, pp. 111-120
It is well known that deep neural networks (DNNs) are vulnerable to adversarial attacks, which are implemented by adding crafted perturbations onto benign examp les. Min-max robust optimization based adversarial training can provide a notion of security against adversarial attacks. However, adversarial robustness requir es a significantly larger capacity of the network than that for the natural trai ning with only benign examples. This paper proposes a framework of concurrent ad versarial training and weight pruning that enables model compression while still preserving the adversarial robustness and essentially tackles the dilemma of ad versarial training. Furthermore, this work studies two hypotheses about weight p runing in the conventional setting and finds that weight pruning is essential fo r reducing the network model size in the adversarial setting; training a small m odel from scratch even with inherited initialization from the large model cannot achieve neither adversarial robustness nor high standard accuracy. Code is avai lable at https://github.com/yeshaokai/Robustness-Aware-Pruning-ADMM.
****

On the Design of Black-Box Adversarial Examples by Leveraging Gradient-Free Opti mization and Operator Splitting Method
Pu Zhao, Sijia Liu, Pin-Yu Chen, Nghia Hoang, Kaidi Xu, Bhavya Kailkhura, Xue Lin; Proceedings of the IEEE/CVF International Conference on Computer Vision

(ICCV), 2019, pp. 121-130
Robust machine learning is currently one of the most prominent topics which could potentially help shaping a future of advanced AI platforms that not only perform well in average cases but also in worst cases or adverse situations. Despite the long-term vision, however, existing studies on black-box adversarial attacks are still restricted to very specific settings of threat models (e.g., single distortion metric and restrictive assumption on target model's feedback to queries) and/or suffer from prohibitively high query complexity. To push for further advances in this field, we introduce a general framework based on an operator splitting method, the alternating direction method of multipliers (ADMM) to devise efficient, robust black-box attacks that work with various distortion metrics and feedback settings without incurring high query complexity. Due to the black-box nature of the threat model, the proposed ADMM solution framework is integrated with zeroth-order (ZO) optimization and Bayesian optimization (BO), and thus is applicable to the gradient-free regime. This results in two new black-box adversarial attack generation methods, ZO-ADMM and BO-ADMM. Our empirical evaluations on image classification datasets show that our proposed approaches have much lower function query complexities compared to state-of-the-art attack methods, but achieve very competitive attack success rates.
****

DewarpNet: Single-Image Document Unwarping With Stacked 3D and 2D Regression Networks
Sagnik Das, Ke Ma, Zhixin Shu, Dimitris Samaras, Roy Shilkrot; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 131-140
Capturing document images with hand-held devices in unstructured environments is a common practice nowadays. However, "casual" photos of documents are usually unsuitable for automatic information extraction, mainly due to physical distortion of the document paper, as well as various camera positions and illumination conditions. In this work, we propose DewarpNet, a deep-learning approach for document image unwarping from a single image. Our insight is that the 3D geometry of the document not only determines the warping of its texture but also causes the illumination effects. Therefore, our novelty resides on the explicit modeling of 3D shape for document paper in an end-to-end pipeline. Also, we contribute the largest and most comprehensive dataset for document image unwarping to date - Doc3D. This dataset features multiple ground-truth annotations, including 3D shape, surface normals, UV map, albedo image, etc. Training with Doc3D, we demonstrate state-of-the-art performance for DewarpNet with extensive qualitative and quantitative evaluations. Our network also significantly improves OCR performance on captured document images, decreasing character error rate by 42% on average. Both the code and the dataset are released.
****

Learning Robust Facial Landmark Detection via Hierarchical Structured Ensemble
Xu Zou, Sheng Zhong, Luxin Yan, Xiangyun Zhao, Jiahuan Zhou, Ying Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 141-150
Heatmap regression-based models have significantly advanced the progress of facial landmark detection. However, the lack of structural constraints always generates inaccurate heatmaps resulting in poor landmark detection performance. While hierarchical structure modeling methods have been proposed to tackle this issue, they all heavily rely on manually designed tree structures. The designed hierarchical structure is likely to be completely corrupted due to the missing or inaccurate prediction of landmarks. To the best of our knowledge, in the context of deep learning, no work before has investigated how to automatically model proper structures for facial landmarks, by discovering their inherent relations. In this paper, we propose a novel Hierarchical Structured Landmark Ensemble (HSLE) model for learning robust facial landmark detection, by using it as the structural constraints. Different from existing approaches of manually designing structures, our proposed HSLE model is constructed automatically via discovering the most robust patterns so HSLE has the ability to robustly depict both local and holis

tic landmark structures simultaneously. Our proposed HSLE can be readily plugged into any existing facial landmark detection baselines for further performance improvement. Extensive experimental results demonstrate our approach significantly outperforms the baseline by a large margin to achieve a state-of-the-art performance.
****

Remote Heart Rate Measurement From Highly Compressed Facial Videos: An End-to-End Deep Learning Solution With Video Enhancement
Zitong Yu, Wei Peng, Xiaobai Li, Xiaopeng Hong, Guoying Zhao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 151-160
Remote photoplethysmography (rPPG), which aims at measuring heart activities without any contact, has great potential in many applications (e.g., remote healthcare). Existing rPPG approaches rely on analyzing very fine details of facial videos, which are prone to be affected by video compression. Here we propose a two-stage, end-to-end method using hidden rPPG information enhancement and attention networks, which is the first attempt to counter video compression loss and recover rPPG signals from highly compressed videos. The method includes two parts: 1) a Spatio-Temporal Video Enhancement Network (STVEN) for video enhancement, and 2) an rPPG network (rPPGNet) for rPPG signal recovery. The rPPGNet can work on its own for robust rPPG measurement, and the STVEN network can be added and jointly trained to further boost the performance especially on highly compressed videos. Comprehensive experiments are performed on two benchmark datasets to show that, 1) the proposed method not only achieves superior performance on compressed videos with high-quality videos pair, 2) it also generalizes well on novel data with only compressed videos available, which implies the promising potential for real-world applications.
****

Face-to-Parameter Translation for Game Character Auto-Creation
Tianyang Shi, Yi Yuan, Changjie Fan, Zhengxia Zou, Zhenwei Shi, Yong Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 161-170
Character customization system is an important component in Role-Playing Games (RPGs), where players are allowed to edit the facial appearance of their in-game characters with their own preferences rather than using default templates. This paper proposes a method for automatically creating in-game characters of players according to an input face photo. We formulate the above "artistic creation" process under a facial similarity measurement and parameter searching paradigm by solving an optimization problem over a large set of physically meaningful facial parameters. To effectively minimize the distance between the created face and the real one, two loss functions, i.e. a "discriminative loss" and a "facial content loss", are specifically designed. As the rendering process of a game engine is not differentiable, a generative network is further introduced as an "imitator" to imitate the physical behavior of the game engine so that the proposed method can be implemented under a neural style transfer framework and the parameters can be optimized by gradient descent. Experimental results demonstrate that our method achieves a high degree of generation similarity between the input face photo and the created in-game character in terms of both global appearance and local details. Our method has been deployed in a new game last year and has now been used by players over 1 million times.
****

Visual Deprojection: Probabilistic Recovery of Collapsed Dimensions
Guha Balakrishnan, Adrian V. Dalca, Amy Zhao, John V. Guttag, Fredo Durand, William T. Freeman; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 171-180
We introduce visual deprojection: the task of recovering an image or video that has been collapsed along a dimension. Projections arise in various contexts, such as long-exposure photography, where a dynamic scene is collapsed in time to produce a motion-blurred image, and corner cameras, where reflected light from a scene is collapsed along a spatial dimension because of an edge occluder to yield

a 1D video. Deprojection is ill-posed-- often there are many plausible solutions for a given input. We first propose a probabilistic model capturing the ambiguity of the task. We then present a variational inference strategy using convolutional neural networks as functional approximators. Sampling from the inference network at test time yields plausible candidates from the distribution of original signals that are consistent with a given input projection. We evaluate the method on several datasets for both spatial and temporal deprojection tasks. We first demonstrate the method can recover human gait videos and face images from spatial projections, and then show that it can recover videos of moving digits from dramatically motion-blurred images obtained via temporal projection.
****

StructureFlow: Image Inpainting via Structure-Aware Appearance Flow
Yurui Ren, Xiaoming Yu, Ruonan Zhang, Thomas H. Li, Shan Liu, Ge Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 181-190
Image inpainting techniques have shown significant improvements by using deep neural networks recently. However, most of them may either fail to reconstruct reasonable structures or restore fine-grained textures. In order to solve this problem, in this paper, we propose a two-stage model which splits the inpainting task into two parts: structure reconstruction and texture generation. In the first stage, edge-preserved smooth images are employed to train a structure reconstructor which completes the missing structures of the inputs. In the second stage, based on the reconstructed structures, a texture generator using appearance flow is designed to yield image details. Experiments on multiple publicly available datasets show the superior performance of the proposed network.
****

Learning Fixed Points in Generative Adversarial Networks: From Image-to-Image Translation to Disease Detection and Localization
Md Mahfuzur Rahman Siddiquee, Zongwei Zhou, Nima Tajbakhsh, Ruibin Feng, Michael B. Gotway, Yoshua Bengio, Jianming Liang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 191-200
Generative adversarial networks (GANs) have ushered in a revolution in image-to-image translation. The development and proliferation of GANs raises an interesting question: can we train a GAN to remove an object, if present, from an image while otherwise preserving the image? Specifically, can a GAN "virtually heal" anyone by turning his medical image, with an unknown health status (diseased or healthy), into a healthy one, so that diseased regions could be revealed by subtracting those two images? Such a task requires a GAN to identify a minimal subset of target pixels for domain translation, an ability that we call fixed-point translation, which no GAN is equipped with yet. Therefore, we propose a new GAN, called Fixed-Point GAN, trained by (1) supervising same-domain translation through a conditional identity loss, and (2) regularizing cross-domain translation through revised adversarial, domain classification, and cycle consistency loss. Based on fixed-point translation, we further derive a novel framework for disease detection and localization using only image-level annotation. Qualitative and quantitative evaluations demonstrate that the proposed method outperforms the state of the art in multi-domain image-to-image translation and that it surpasses predominant weakly-supervised localization methods in both disease detection and localization. Implementation is available at https://github.com/jlianglab/Fixed-Point-GAN.
****

Generative Adversarial Training for Weakly Supervised Cloud Matting
Zhengxia Zou, Wenyuan Li, Tianyang Shi, Zhenwei Shi, Jieping Ye; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 201-210
The detection and removal of cloud in remote sensing images are essential for earth observation applications. Most previous methods consider cloud detection as a pixel-wise semantic segmentation process (cloud v.s. background), which inevitably leads to a category-ambiguity problem when dealing with semi-transparent clouds. We re-examine the cloud detection under a totally different point of view,

i.e. to formulate it as a mixed energy separation process between foreground and background images, which can be equivalently implemented under an image matting paradigm with a clear physical significance. We further propose a generative adversarial framework where the training of our model neither requires any pixel-wise ground truth reference nor any additional user interactions. Our model consists of three networks, a cloud generator G, a cloud discriminator D, and a cloud matting network F, where G and D aim to generate realistic and physically meaningful cloud images by adversarial training, and F learns to predict the cloud reflectance and attenuation. Experimental results on a global set of satellite images demonstrate that our method, without ever using any pixel-wise ground truth during training, achieves comparable and even higher accuracy over other fully supervised methods, including some recent popular cloud detectors and some well-known semantic segmentation frameworks.

****

PAMTRI: Pose-Aware Multi-Task Learning for Vehicle Re-Identification Using Highly Randomized Synthetic Data

Zheng Tang, Milind Naphade, Stan Birchfield, Jonathan Tremblay, William Hodge, Ratnesh Kumar, Shuo Wang, Xiaodong Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 211-220

In comparison with person re-identification (ReID), which has been widely studied in the research community, vehicle ReID has received less attention. Vehicle ReID is challenging due to 1) high intra-class variability (caused by the dependency of shape and appearance on viewpoint), and 2) small inter-class variability (caused by the similarity in shape and appearance between vehicles produced by different manufacturers). To address these challenges, we propose a Pose-Aware Multi-Task Re-Identification (PAMTRI) framework. This approach includes two innovations compared with previous methods. First, it overcomes viewpoint-dependency by explicitly reasoning about vehicle pose and shape via keypoints, heatmaps and segments from pose estimation. Second, it jointly classifies semantic vehicle attributes (colors and types) while performing ReID, through multi-task learning with the embedded pose representations. Since manually labeling images with detailed pose and attribute information is prohibitive, we create a large-scale highly randomized synthetic dataset with automatically annotated vehicle attributes for training. Extensive experiments validate the effectiveness of each proposed component, showing that PAMTRI achieves significant improvement over state-of-the-art on two mainstream vehicle ReID benchmarks: VeRi and CityFlow-ReID.

****

Generative Adversarial Networks for Extreme Learned Image Compression

Eirikur Agustsson, Michael Tschannen, Fabian Mentzer, Radu Timofte, Luc Van Gool; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 221-231

We present a learned image compression system based on GANs, operating at extremely low bitrates. Our proposed framework combines an encoder, decoder/generator and a multi-scale discriminator, which we train jointly for a generative learned compression objective. The model synthesizes details it cannot afford to store, obtaining visually pleasing results at bitrates where previous methods fail and show strong artifacts. Furthermore, if a semantic label map of the original image is available, our method can fully synthesize unimportant regions in the decoded image such as streets and trees from the label map, proportionally reducing the storage cost. A user study confirms that for low bitrates, our approach is preferred to state-of-the-art methods, even when they use more than double the bits.

****

Instance-Guided Context Rendering for Cross-Domain Person Re-Identification

Yanbei Chen, Xiatian Zhu, Shaogang Gong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 232-242

Existing person re-identification (re-id) methods mostly assume the availability of large-scale identity labels for model learning in any target domain deployment. This greatly limits their scalability in practice. To tackle this limitation, we propose a novel Instance-Guided Context Rendering scheme, which transfers t

he source person identities into diverse target domain contexts to enable superv ised re-id model learning in the unlabelled target domain. Unlike previous image synthesis methods that transform the source person images into limited fixed ta rget styles, our approach produces more visually plausible, and diverse syntheti c training data. Specifically, we formulate a dual conditional generative advers arial network that augments each source person image with rich contextual variat ions. To explicitly achieve diverse rendering effects, we leverage abundant unla belled target instances as contextual guidance for image generation. Extensive e xperiments on Market-1501, DukeMTMC-reID and CUHK03 benchmarks show that the re-id performance can be significantly improved when using our synthetic data in cr oss-domain re-id model learning.

****

What Else Can Fool Deep Learning? Addressing Color Constancy Errors on Deep Neur al Network Performance

Mahmoud Afifi, Michael S. Brown; Proceedings of the IEEE/CVF International Conf erence on Computer Vision (ICCV), 2019, pp. 243-252

There is active research targeting local image manipulations that can fool deep neural networks (DNNs) into producing incorrect results. This paper examines a t ype of global image manipulation that can produce similar adverse effects. Speci fically, we explore how strong color casts caused by incorrectly applied computa tional color constancy - referred to as white balance (WB) in photography - nega tively impact the performance of DNNs targeting image segmentation and classific ation. In addition, we discuss how existing image augmentation methods used to i mprove the robustness of DNNs are not well suited for modeling WB errors. To add ress this problem, a novel augmentation method is proposed that can emulate accu rate color constancy degradation. We also explore pre-processing training and te sting images with a recent WB correction algorithm to reduce the effects of inco rrectly white-balanced images. We examine both augmentation and pre-processing s trategies on different datasets and demonstrate notable improvements on the CIFA R-10, CIFAR-100, and ADE20K datasets.

****

Beyond Cartesian Representations for Local Descriptors

Patrick Ebel, Anastasiia Mishchuk, Kwang Moo Yi, Pascal Fua, Eduard Trulls; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 253-262

The dominant approach for learning local patch descriptors relies on small image regions whose scale must be properly estimated a priori by a keypoint detector. In other words, if two patches are not in correspondence, their descriptors wil l not match. A strategy often used to alleviate this problem is to "pool" the pi xel-wise features over log-polar regions, rather than regularly spaced ones. By contrast, we propose to extract the "support region" directly with a log-polar s ampling scheme. We show that this provides us with a better representation by si multaneously oversampling the immediate neighbourhood of the point and undersamp ling regions far away from it. We demonstrate that this representation is partic ularly amenable to learning descriptors with deep networks. Our models can match descriptors across a much wider range of scales than was possible before, and a lso leverage much larger support regions without suffering from occlusions. We r eport state-of-the-art results on three different datasets

****

Distilling Knowledge From a Deep Pose Regressor Network

Muhamad Risqi U. Saputra, Pedro P. B. de Gusmao, Yasin Almalioglu, Andrew Mar kham, Niki Trigoni; Proceedings of the IEEE/CVF International Conference on Com puter Vision (ICCV), 2019, pp. 263-272

This paper presents a novel method to distill knowledge from a deep pose regress or network for efficient Visual Odometry (VO). Standard distillation relies on " dark knowledge" for successful knowledge transfer. As this knowledge is not avai lable in pose regression and the teacher prediction is not always accurate, we p ropose to emphasize the knowledge transfer only when we trust the teacher. We ac hieve this by using teacher loss as a confidence score which places variable rel ative importance on the teacher prediction. We inject this confidence score to t

he main training task via Attentive Imitation Loss (AIL) and when learning the i
ntermediate representation of the teacher through Attentive Hint Training (AHT)
approach. To the best of our knowledge, this is the first work which successfull
y distill the knowledge from a deep pose regression network. Our evaluation on t
he KITTI and Malaga dataset shows that we can keep the student prediction close
to the teacher with up to 92.95% parameter reduction and 2.12x faster in computa
tion time.
****

Instance-Level Future Motion Estimation in a Single Image Based on Ordinal Regre
ssion

Kyung-Rae Kim, Whan Choi, Yeong Jun Koh, Seong-Gyun Jeong, Chang-Su Kim; Pro
ceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 201
9, pp. 273-282

A novel algorithm to estimate instance-level future motion in a single image is
proposed in this paper. We first represent the future motion of an instance with
its direction, speed, and action classes. Then, we develop a deep neural networ
k that exploits different levels of semantic information to perform the future m
otion estimation. For effective future motion classification, we adopt ordinal r
egression. Especially, we develop the cyclic ordinal regression scheme using bin
ary classifiers. Experiments demonstrate that the proposed algorithm provides re
liable performance and thus can be used effectively for vision applications, inc
luding single and multi object tracking. Furthermore, we release the future moti
on (FM) dataset, collected from diverse sources and annotated manually, as a ben
chmark for single-image future motion estimation.
****

Vision-Infused Deep Audio Inpainting

Hang Zhou, Ziwei Liu, Xudong Xu, Ping Luo, Xiaogang Wang; Proceedings of the
IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 283-292

Multi-modality perception is essential to develop interactive intelligence. In t
his work, we consider a new task of visual information-infused audio inpainting,
i.e., synthesizing missing audio segments that correspond to their accompanying
videos. We identify two key aspects for a successful inpainter: (1) It is desir
able to operate on spectrograms instead of raw audios. Recent advances in deep s
emantic image inpainting could be leveraged to go beyond the limitations of trad
itional audio inpainting. (2) To synthesize visually indicated audio, a visual-a
udio joint feature space needs to be learned with synchronization of audio and v
ideo. To facilitate a large-scale study, we collect a new multi-modality instrum
ent-playing dataset called MUSIC-Extra-Solo (MUSICES) by enriching MUSIC dataset
. Extensive experiments demonstrate that our framework is capable of inpainting
realistic and varying audio segments with or without visual contexts. More impor
tantly, our synthesized audio segments are coherent with their video counterpart
s, showing the effectiveness of our proposed Vision-Infused Audio Inpainter (VIA
I).
****

HAWQ: Hessian AWare Quantization of Neural Networks With Mixed-Precision

Zhen Dong, Zhewei Yao, Amir Gholami, Michael W. Mahoney, Kurt Keutzer; Proce
edings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019,
pp. 293-302

Model size and inference speed/power have become a major challenge in the deploy
ment of neural networks for many applications. A promising approach to address t
hese problems is quantization. However, uniformly quantizing a model to ultra-lo
w precision leads to significant accuracy degradation. A novel solution for this
is to use mixed-precision quantization, as some parts of the network may allow
lower precision as compared to other layers. However, there is no systematic way
to determine the precision of different layers. A brute force approach is not f
easible for deep networks, as the search space for mixed-precision is exponentia
l in the number of layers. Another challenge is a similar factorial complexity f
or determining block-wise fine-tuning order when quantizing the model to a targe
t precision. Here, we introduce Hessian AWare Quantization (HAWQ), a novel secon
d-order quantization method to address these problems. HAWQ allows for the autom

atic selection of the relative quantization precision of each layer, based on th e layer's Hessian spectrum. Moreover, HAWQ provides a deterministic fine-tuning order for quantizing layers. We show the results of our method on Cifar-10 using ResNet20, and on ImageNet using Inception-V3, ResNet50 and SqueezeNext models. Comparing HAWQ with state-of-the-art shows that we can achieve similar/better ac curacy with 8x activation compression ratio on ResNet20, as compared to DNAS, an d up to 1% higher accuracy with up to 14% smaller models on ResNet50 and Incepti on-V3, compared to recently proposed methods of RVQuant and HAQ. Furthermore, we show that we can quantize SqueezeNext to just 1MB model size while achieving ab ove 68% top1 accuracy on ImageNet.
****

Evaluating Robustness of Deep Image Super-Resolution Against Adversarial Attacks
Jun-Ho Choi, Huan Zhang, Jun-Hyuk Kim, Cho-Jui Hsieh, Jong-Seok Lee; Proceed ings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, p p. 303-311
Single-image super-resolution aims to generate a high-resolution version of a lo w-resolution image, which serves as an essential component in many image process ing applications. This paper investigates the robustness of deep learning-based super-resolution methods against adversarial attacks, which can significantly de teriorate the super-resolved images without noticeable distortion in the attacke d low-resolution images. It is demonstrated that state-of-the-art deep super-res olution methods are highly vulnerable to adversarial attacks. Different levels o f robustness of different methods are analyzed theoretically and experimentally. We also present analysis on transferability of attacks, and feasibility of targ eted attacks and universal attacks.
****

Overcoming Catastrophic Forgetting With Unlabeled Data in the Wild
Kibok Lee, Kimin Lee, Jinwoo Shin, Honglak Lee; Proceedings of the IEEE/CVF I nternational Conference on Computer Vision (ICCV), 2019, pp. 312-321
Lifelong learning with deep neural networks is well-known to suffer from catastr ophic forgetting: the performance on previous tasks drastically degrades when le arning a new task. To alleviate this effect, we propose to leverage a large stre am of unlabeled data easily obtainable in the wild. In particular, we design a n ovel class-incremental learning scheme with (a) a new distillation loss, termed global distillation, (b) a learning strategy to avoid overfitting to the most re cent task, and (c) a confidence-based sampling method to effectively leverage un labeled external data. Our experimental results on various datasets, including C IFAR and ImageNet, demonstrate the superiority of the proposed methods over prio r methods, particularly when a stream of unlabeled data is accessible: our metho d shows up to 15.8% higher accuracy and 46.5% less forgetting compared to the st ate-of-the-art method. The code is available at https://github.com/kibok90/iccv2 019-inc.
****

Symmetric Cross Entropy for Robust Learning With Noisy Labels
Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, James Bailey; Pro ceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 201 9, pp. 322-330
Training accurate deep neural networks (DNNs) in the presence of noisy labels is an important and challenging task. Though a number of approaches have been prop osed for learning with noisy labels, many open issues remain. In this paper, we show that DNN learning with Cross Entropy (CE) exhibits overfitting to noisy lab els on some classes ("easy" classes), but more surprisingly, it also suffers fro m significant under learning on some other classes ("hard" classes). Intuitively , CE requires an extra term to facilitate learning of hard classes, and more imp ortantly, this term should be noise tolerant, so as to avoid overfitting to nois y labels. Inspired by the symmetric KL-divergence, we propose the approach of Sy mmetric cross entropy Learning (SL), boosting CE symmetrically with a noise robu st counterpart Reverse Cross Entropy (RCE). Our proposed SL approach simultaneou sly addresses both the under learning and overfitting problem of CE in the prese nce of noisy labels. We provide a theoretical analysis of SL and also empiricall

y show, on a range of benchmark and real-world datasets, that SL outperforms sta te-of-the-art methods. We also show that SL can be easily incorporated into exis ting methods in order to further enhance their performance.
****

Few-Shot Learning With Embedded Class Models and Shot-Free Meta Training
Avinash Ravichandran, Rahul Bhotika, Stefano Soatto; Proceedings of the IEEE/C VF International Conference on Computer Vision (ICCV), 2019, pp. 331-339
We propose a method for learning embeddings for few-shot learning that is suitab le for use with any number of shots (shot-free). Rather than fixing the class pr ototypes to be the Euclidean average of sample embeddings, we allow them to live in a higher-dimensional space (embedded class models) and learn the prototypes along with the model parameters. The class representation function is defined im plicitly, which allows us to deal with a variable number of shots per class with a simple constant-size architecture. The class embedding encompasses metric lea rning, that facilitates adding new classes without crowding the class representa tion space. Despite being general and not tuned to the benchmark, our approach a chieves state-of-the-art performance on the standard few-shot benchmark datasets .
****

Dual Directed Capsule Network for Very Low Resolution Image Recognition
Maneet Singh, Shruti Nagpal, Richa Singh, Mayank Vatsa; Proceedings of the IE EE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 340-349
Very low resolution (VLR) image recognition corresponds to classifying images wi th resolution 16x16 or less. Though it has widespread applicability when objects are captured at a very large stand-off distance (e.g. surveillance scenario) or from wide angle mobile cameras, it has received limited attention. This researc h presents a novel Dual Directed Capsule Network model, termed as DirectCapsNet, for addressing VLR digit and face recognition. The proposed architecture utiliz es a combination of capsule and convolutional layers for learning an effective V LR recognition model. The architecture also incorporates two novel loss function s: (i) the proposed HR-anchor loss and (ii) the proposed targeted reconstruction loss, in order to overcome the challenges of limited information content in VLR images. The proposed losses use high resolution images as auxiliary data during training to "direct" discriminative feature learning. Multiple experiments for VLR digit classification and VLR face recognition are performed along with compa risons with state-of-the-art algorithms. The proposed DirectCapsNet consistently showcases state-of-the-art results; for example, on the UCCS face database, it shows over 95% face recognition accuracy when 16x16 images are matched with 80x8 0 images.
****

Recognizing Part Attributes With Insufficient Data
Xiangyun Zhao, Yi Yang, Feng Zhou, Xiao Tan, Yuchen Yuan, Yingze Bao, Ying Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (IC CV), 2019, pp. 350-360
Recognizing the attributes of objects and their parts is central to many compute r vision applications. Although great progress has been made to apply object-lev el recognition, recognizing the attributes of parts remains less applicable sinc e the training data for part attributes recognition is usually scarce especially for internet-scale applications. Furthermore, most existing part attribute reco gnition methods rely on the part annotations which are more expensive to obtain. In order to solve the data insufficiency problem and get rid of dependence on t he part annotation, we introduce a novel Concept Sharing Network (CSN) for part attribute recognition. A great advantage of CSN is its capability of recognizing the part attribute (a combination of part location and appearance pattern) that has insufficient or zero training data, by learning the part location and appea rance pattern respectively from the training data that usually mix them in a sin gle label. Extensive experiments on CUB, Celeb A, and a newly proposed human att ribute dataset demonstrate the effectiveness of CSN and its advantages over othe r methods, especially for the attributes with few training samples. Further expe riments show that CSN can also perform zero-shot part attribute recognition.

****
USIP: Unsupervised Stable Interest Point Detection From 3D Point Clouds
Jiaxin Li, Gim Hee Lee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 361-370
In this paper, we propose the USIP detector: an Unsupervised Stable Interest Point detector that can detect highly repeatable and accurately localized keypoints from 3D point clouds under arbitrary transformations without the need for any ground truth training data. Our USIP detector consists of a feature proposal network that learns stable keypoints from input 3D point clouds and their respective transformed pairs from randomly generated transformations. We provide degeneracy analysis and suggest solutions to prevent it. We encourage high repeatability and accurate localization of the keypoints with a probabilistic chamfer loss that minimizes the distances between the detected keypoints from the training point cloud pairs. Extensive experimental results of repeatability tests on several simulated and real-world 3D point cloud datasets from Lidar, RGB-D and CAD models show that our USIP detector significantly outperforms existing hand-crafted and deep learning-based 3D keypoint detectors. Our code is available at the project website. https://github.com/lijx10/USIP
****
Mixed High-Order Attention Network for Person Re-Identification
Binghui Chen, Weihong Deng, Jiani Hu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 371-381
Attention has become more attractive in person re-identification (ReID) as it is capable of biasing the allocation of available resources towards the most informative parts of an input signal. However, state-of-the-art works concentrate only on coarse or first-order attention design, e.g. spatial and channels attention, while rarely exploring higher-order attention mechanism. We take a step towards addressing this problem. In this paper, we first propose the High-Order Attention (HOA) module to model and utilize the complex and high-order statistics information in attention mechanism, so as to capture the subtle differences among pedestrians and to produce the discriminative attention proposals. Then, rethinking person ReID as a zero-shot learning problem, we propose the Mixed High-Order Attention Network (MHN) to further enhance the discrimination and richness of attention knowledge in an explicit manner. Extensive experiments have been conducted to validate the superiority of our MHN for person ReID over a wide variety of state-of-the-art methods on three large-scale datasets, including Market-1501, DukeMTMC-ReID and CUHK03-NP. Code is available at http://www.bhchen.cn.
****
Budget-Aware Adapters for Multi-Domain Learning
Rodrigo Berriel, Stephane Lathuillere, Moin Nabi, Tassilo Klein, Thiago Oliveira-Santos, Nicu Sebe, Elisa Ricci; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 382-391
Multi-Domain Learning (MDL) refers to the problem of learning a set of models derived from a common deep architecture, each one specialized to perform a task in a certain domain (e.g., photos, sketches, paintings). This paper tackles MDL with a particular interest in obtaining domain-specific models with an adjustable budget in terms of the number of network parameters and computational complexity. Our intuition is that, as in real applications the number of domains and tasks can be very large, an effective MDL approach should not only focus on accuracy but also on having as few parameters as possible. To implement this idea we derive specialized deep models for each domain by adapting a pre-trained architecture but, differently from other methods, we propose a novel strategy to automatically adjust the computational complexity of the network. To this aim, we introduce Budget-Aware Adapters that select the most relevant feature channels to better handle data from a novel domain. Some constraints on the number of active switches are imposed in order to obtain a network respecting the desired complexity budget. Experimentally, we show that our approach leads to recognition accuracy competitive with state-of-the-art approaches but with much lighter networks both in terms of storage and computation.
****

Compact Trilinear Interaction for Visual Question Answering
Tuong Do, Thanh-Toan Do, Huy Tran, Erman Tjiputra, Quang D. Tran; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 392-401
In Visual Question Answering (VQA), answers have a great correlation with question meaning and visual contents. Thus, to selectively utilize image, question and answer information, we propose a novel trilinear interaction model which simultaneously learns high level associations between these three inputs. In addition, to overcome the interaction complexity, we introduce a multimodal tensor-based PARALIND decomposition which efficiently parameterizes trilinear teraction between the three inputs. Moreover, knowledge distillation is first time applied in Free-form Opened-ended VQA. It is not only for reducing the computational cost and required memory but also for transferring knowledge from trilinear interaction model to bilinear interaction model. The extensive experiments on benchmarking datasets TDIUC, VQA-2.0, and Visual7W show that the proposed compact trilinear interaction model achieves state-of-the-art results when using a single model on all three datasets.
****

Towards Latent Attribute Discovery From Triplet Similarities
Ishan Nigam, Pavel Tokmakov, Deva Ramanan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 402-410
This paper addresses the task of learning latent attributes from triplet similarity comparisons. Consider, for instance, the three shoes in Fig. 1(a). They can be compared according to color, comfort, size, or shape resulting in different rankings. Most approaches for embedding learning either make a simplifying assumption - that all inputs are comparable under a single criterion, or require expensive attribute supervision. We introduce Latent Similarity Networks (LSNs): a simple and effective technique to discover the underlying latent notions of similarity in data without any explicit attribute supervision. LSNs can be trained with standard triplet supervision and learn several latent embeddings that can be used to compare images under multiple notions of similarity. LSNs achieve state-of-the-art performance on UT-Zappos-50k Shoes and Celeb-A Faces datasets and also demonstrate the ability to uncover meaningful latent attributes.
****

GeoStyle: Discovering Fashion Trends and Events
Utkarsh Mall, Kevin Matzen, Bharath Hariharan, Noah Snavely, Kavita Bala; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 411-420
Understanding fashion styles and trends is of great potential interest to retailers and consumers alike. The photos people upload to social media are a historical and public data source of how people dress across the world and at different times. While we now have tools to automatically recognize the clothing and style attributes of what people are wearing in these photographs, we lack the ability to analyze spatial and temporal trends in these attributes or make predictions about the future. In this paper we address this need by providing an automatic framework that analyzes large corpora of street imagery to (a) discover and forecast long-term trends of various fashion attributes as well as automatically discovered styles, and (b) identify spatio-temporally localized events that affect what people wear. We show that our framework makes long term trend forecasts that are > 20% more accurate than prior art, and identifies hundreds of socially meaningful events that impact fashion across the globe.
****

Towards Adversarially Robust Object Detection
Haichao Zhang, Jianyu Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 421-430
Object detection is an important vision task and has emerged as an indispensable component in many vision system, rendering its robustness as an increasingly important performance factor for practical applications. While object detection models have been demonstrated to be vulnerable against adversarial attacks by many recent works, very few efforts have been devoted to improving their robustness.

In this work, we take an initial attempt towards this direction. We first revisit and systematically analyze object detectors and many recently developed attacks from the perspective of model robustness. We then present a multi-task learning perspective of object detection and identify an asymmetric role of task losses. We further develop an adversarial training approach which can leverage the multiple sources of attacks for improving the robustness of detection models. Extensive experiments on PASCAL-VOC and MS-COCO verified the effectiveness of the proposed approach.
****

Automatic and Robust Skull Registration Based on Discrete Uniformization
Junli Zhao, Xin Qi, Chengfeng Wen, Na Lei, Xianfeng Gu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 431-440
Skull registration plays a fundamental role in forensic science and is crucial for craniofacial reconstruction. The complicated topology, lack of anatomical features, and low quality reconstructed mesh make skull registration challenging. In this work, we propose an automatic skull registration method based on the discrete uniformization theory, which can handle complicated topologies and is robust to low quality meshes. We apply dynamic Yamabe flow to realize discrete uniformization, which modifies the mesh combinatorial structure during the flow and conformally maps the multiply connected skull surface onto a planar disk with circular holes. The 3D surfaces can be registered by matching their planar images using harmonic maps. This method is rigorous with theoretic guarantee, automatic without user intervention, and robust to low mesh quality. Our experimental results demonstrate the efficiency and efficacy of the method.
****

Few-Shot Image Recognition With Knowledge Transfer
Zhimao Peng, Zechao Li, Junge Zhang, Yan Li, Guo-Jun Qi, Jinhui Tang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 441-449
Human can well recognize images of novel categories just after browsing few examples of these categories. One possible reason is that they have some external discriminative visual information about these categories from their prior knowledge. Inspired from this, we propose a novel Knowledge Transfer Network architecture (KTN) for few-shot image recognition. The proposed KTN model jointly incorporates visual feature learning, knowledge inferring and classifier learning into one unified framework for their optimal compatibility. First, the visual classifiers for novel categories are learned based on the convolutional neural network with the cosine similarity optimization. To fully explore the prior knowledge, a semantic-visual mapping network is then developed to conduct knowledge inference, which enables to infer the classifiers for novel categories from base categories. Finally, we design an adaptive fusion scheme to infer the desired classifiers by effectively integrating the above knowledge and visual information. Extensive experiments are conducted on two widely-used Mini-ImageNet and ImageNet Few-Shot benchmarks to evaluate the effectiveness of the proposed method. The results compared with the state-of-the-art approaches show the encouraging performance of the proposed method, especially on 1-shot and 2-shot tasks.
****

Fine-Grained Action Retrieval Through Multiple Parts-of-Speech Embeddings
Michael Wray, Diane Larlus, Gabriela Csurka, Dima Damen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 450-459
We address the problem of cross-modal fine-grained action retrieval between text and video. Cross-modal retrieval is commonly achieved through learning a shared embedding space, that can indifferently embed modalities. In this paper, we propose to enrich the embedding by disentangling parts-of-speech (PoS) in the accompanying captions. We build a separate multi-modal embedding space for each PoS tag. The outputs of multiple PoS embeddings are then used as input to an integrated multi-modal space, where we perform action retrieval. All embeddings are trained jointly through a combination of PoS-aware and PoS-agnostic losses. Our proposal enables learning specialised embedding spaces that offer multiple views of the same embedded entities. We report the first retrieval results on fine-graine

d actions for the large-scale EPIC dataset, in a generalised zero-shot setting. Results show the advantage of our approach for both video-to-text and text-to-video action retrieval. We also demonstrate the benefit of disentangling the PoS for the generic task of cross-modal video retrieval on the MSR-VTT dataset.
****

Vehicle Re-Identification in Aerial Imagery: Dataset and Approach
Peng Wang, Bingliang Jiao, Lu Yang, Yifei Yang, Shizhou Zhang, Wei Wei, Yanning Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 460-469
In this work, we construct a large-scale dataset for vehicle re-identification (ReID), which contains 137k images of 13k vehicle instances captured by UAV-mounted cameras. To our knowledge, it is the largest UAV-based vehicle ReID dataset. To increase intra-class variation, each vehicle is captured by at least two UAVs at different locations, with diverse view-angles and flight-altitudes. We manually label a variety of vehicle attributes, including vehicle type, color, skylight, bumper, spare tire and luggage rack. Furthermore, for each vehicle image, the annotator is also required to mark the discriminative parts that helps them to distinguish this particular vehicle from others. Besides the dataset, we also design a specific vehicle ReID algorithm to make full use of the rich annotation information. It is capable of explicitly detecting discriminative parts for each specific vehicle and significantly outperforming the evaluated baselines and state-of-the-art vehicle ReID approaches.
****

Bridging the Domain Gap for Ground-to-Aerial Image Matching
Krishna Regmi, Mubarak Shah; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 470-479
The visual entities in cross-view (e.g. ground and aerial) images exhibit drastic domain changes due to the differences in viewpoints each set of images is captured from. Existing state-of-the-art methods address the problem by learning view-invariant images descriptors. We propose a novel method for solving this task by exploiting the gener- ative powers of conditional GANs to synthesize an aerial representation of a ground-level panorama query and use it to minimize the domain gap between the two views. The synthesized image being from the same view as the ref- erence (target) image, helps the network to preserve im- portant cues in aerial images following our Joint Feature Learning approach. We fuse the complementary features from a synthesized aerial image with the original ground- level panorama features to obtain a robust query represen- tation. In addition, we employ multi-scale feature aggre- gation in order to preserve image representations at dif- ferent scales useful for solving this complex task. Experi- mental results show that our proposed approach performs significantly better than the state-of-the-art methods on the challenging CVUSA dataset in terms of top-1 and top-1% retrieval accuracies. Furthermore, we evaluate the gen- eralization of the proposed method for urban landscapes on our newly collected cross-view localization dataset with geo-reference information.
****

A Robust Learning Approach to Domain Adaptive Object Detection
Mehran Khodabandeh, Arash Vahdat, Mani Ranjbar, William G. Macready; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 480-490
Domain shift is unavoidable in real-world applications of object detection. For example, in self-driving cars, the target domain consists of unconstrained road environments which cannot all possibly be observed in training data. Similarly, in surveillance applications sufficiently representative training data may be lacking due to privacy regulations. In this paper, we address the domain adaptation problem from the perspective of robust learning and show that the problem may be formulated as training with noisy labels. We propose a robust object detection framework that is resilient to noise in bounding box class labels, locations and size annotations. To adapt to the domain shift, the model is trained on the target domain using a set of noisy object bounding boxes that are obtained by a detection model trained only in the source domain. We evaluate the accuracy of ou

r approach in various source/target domain pairs and demonstrate that the model significantly improves the state-of-the-art on multiple domain adaptation scenarios on the SIM10K, Cityscapes and KITTI datasets.
****

Graph-Based Object Classification for Neuromorphic Vision Sensing
Yin Bi, Aaron Chadha, Alhabib Abbas, Eirina Bourtsoulatze, Yiannis Andreopoulos; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 491-501
Neuromorphic vision sensing (NVS) devices represent visual information as sequences of asynchronous discrete events (a.k.a., "spikes'") in response to changes in scene reflectance. Unlike conventional active pixel sensing (APS), NVS allows for significantly higher event sampling rates at substantially increased energy efficiency and robustness to illumination changes. However, object classification with NVS streams cannot leverage on state-of-the-art convolutional neural networks (CNNs), since NVS does not produce frame representations. To circumvent this mismatch between sensing and processing with CNNs, we propose a compact graph representation for NVS. We couple this with novel residual graph CNN architectures and show that, when trained on spatio-temporal NVS data for object classification, such residual graph CNNs preserve the spatial and temporal coherence of spike events, while requiring less computation and memory. Finally, to address the absence of large real-world NVS datasets for complex recognition tasks, we present and make available a 100k dataset of NVS recordings of the American sign language letters, acquired with an iniLabs DAVIS240c device under real-world conditions.
****

Gaussian YOLOv3: An Accurate and Fast Object Detector Using Localization Uncertainty for Autonomous Driving
Jiwoong Choi, Dayoung Chun, Hyun Kim, Hyuk-Jae Lee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 502-511
The use of object detection algorithms is becoming increasingly important in autonomous vehicles, and object detection at high accuracy and a fast inference speed is essential for safe autonomous driving. A false positive (FP) from a false localization during autonomous driving can lead to fatal accidents and hinder safe and efficient driving. Therefore, a detection algorithm that can cope with mislocalizations is required in autonomous driving applications. This paper proposes a method for improving the detection accuracy while supporting a real-time operation by modeling the bounding box (bbox) of YOLOv3, which is the most representative of one-stage detectors, with a Gaussian parameter and redesigning the loss function. In addition, this paper proposes a method for predicting the localization uncertainty that indicates the reliability of bbox. By using the predicted localization uncertainty during the detection process, the proposed schemes can significantly reduce the FP and increase the true positive (TP), thereby improving the accuracy. Compared to a conventional YOLOv3, the proposed algorithm, Gaussian YOLOv3, improves the mean average precision (mAP) by 3.09 and 3.5 on the KITTI and Berkeley deep drive (BDD) datasets, respectively. Nevertheless, the proposed algorithm is capable of real-time detection at faster than 42 frames per second (fps) and shows a higher accuracy than previous approaches with a similar fps. Therefore, the proposed algorithm is the most suitable for autonomous driving applications.
****

Sharpen Focus: Learning With Attention Separability and Consistency
Lezi Wang, Ziyan Wu, Srikrishna Karanam, Kuan-Chuan Peng, Rajat Vikram Singh, Bo Liu, Dimitris N. Metaxas; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 512-521
Recent developments in gradient-based attention modeling have seen attention maps emerge as a powerful tool for interpreting convolutional neural networks. Despite good localization for an individual class of interest, these techniques produce attention maps with substantially overlapping responses among different classes, leading to the problem of visual confusion and the need for discriminative attention. In this paper, we address this problem by means of a new framework th

at makes class-discriminative attention a principled part of the learning proces
s. Our key innovations include new learning objectives for attention separabilit
y and cross-layer consistency, which result in improved attention discriminabili
ty and reduced visual confusion. Extensive experiments on image classification b
enchmarks show the effectiveness of our approach in terms of improved classifica
tion accuracy, including CIFAR-100 (+3.33%), Caltech-256 (+1.64%), ImageNet (+0.
92%), CUB-200-2011 (+4.8%) and PASCAL VOC2012 (+5.73%).
****

Learning Semantic-Specific Graph Representation for Multi-Label Image Recognitio
n

Tianshui Chen, Muxin Xu, Xiaolu Hui, Hefeng Wu, Liang Lin; Proceedings of th
e IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 522-531

Recognizing multiple labels of images is a practical and challenging task, and s
ignificant progress has been made by searching semantic-aware regions and modeli
ng label dependency. However, current methods cannot locate the semantic regions
 accurately due to the lack of part-level supervision or semantic guidance. More
over, they cannot fully explore the mutual interactions among the semantic regio
ns and do not explicitly model the label co-occurrence. To address these issues,
 we propose a Semantic-Specific Graph Representation Learning (SSGRL) framework
that consists of two crucial modules: 1) a semantic decoupling module that incor
porates category semantics to guide learning semantic-specific representations a
nd 2) a semantic interaction module that correlates these representations with a
 graph built on the statistical label co-occurrence and explores their interacti
ons via a graph propagation mechanism. Extensive experiments on public benchmark
s show that our SSGRL framework outperforms current state-of-the-art methods by
a sizable margin, e.g. with an mAP improvement of 2.5%, 2.6%, 6.7%, and 3.1% on
the PASCAL VOC 2007 & 2012, Microsoft-COCO and Visual Genome benchmarks, respect
ively. Our codes and models are available at https://github.com/HCPLab-SYSU/SSGR
L.
****

DeceptionNet: Network-Driven Domain Randomization

Sergey Zakharov, Wadim Kehl, Slobodan Ilic; Proceedings of the IEEE/CVF Intern
ational Conference on Computer Vision (ICCV), 2019, pp. 532-541

We present a novel approach to tackle domain adaptation between synthetic and re
al data. Instead, of employing "blind" domain randomization, i.e., augmenting sy
nthetic renderings with random backgrounds or changing illumination and coloriza
tion, we leverage the task network as its own adversarial guide toward useful au
gmentations that maximize the uncertainty of the output. To this end, we design
a min-max optimization scheme where a given task competes against a special dece
ption network to minimize the task error subject to the specific constraints enf
orced by the deceiver. The deception network samples from a family of differenti
able pixel-level perturbations and exploits the task architecture to find the mo
st destructive augmentations. Unlike GAN-based approaches that require unlabeled
 data from the target domain, our method achieves robust mappings that scale wel
l to multiple target distributions from source data alone. We apply our framewor
k to the tasks of digit recognition on enhanced MNIST variants, classification a
nd object pose estimation on the Cropped LineMOD dataset as well as semantic seg
mentation on the Cityscapes dataset and compare it to a number of domain adaptat
ion approaches, thereby demonstrating similar results with superior generalizati
on capabilities.
****

Pose-Guided Feature Alignment for Occluded Person Re-Identification

Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, Yi Yang; Proceedings of the IEEE/C
VF International Conference on Computer Vision (ICCV), 2019, pp. 542-551

Persons are often occluded by various obstacles in person retrieval scenarios. P
revious person re-identification (re-id) methods, either overlook this issue or
resolve it based on an extreme assumption. To alleviate the occlusion problem, w
e propose to detect the occluded regions, and explicitly exclude those regions d
uring feature generation and matching. In this paper, we introduce a novel metho
d named Pose-Guided Feature Alignment (PGFA), exploiting pose landmarks to disen

tangle the useful information from the occlusion noise. During the feature const
ructing stage, our method utilizes human landmarks to generate attention maps. T
he generated attention maps indicate if a specific body part is occluded and gui
de our model to attend to the non-occluded regions. During matching, we explicit
ly partition the global feature into parts and use the pose landmarks to indicat
e which partial features belonging to the target person. Only the visible region
s are utilized for the retrieval. Besides, we construct a large-scale dataset fo
r the Occluded Person Re-ID problem, namely Occluded-DukeMTMC, which is by far t
he largest dataset for the Occlusion Person Re-ID. Extensive experiments are con
ducted on our constructed occluded re-id dataset, two partial re-id datasets, an
d two commonly used holistic re-id datasets. Our method largely outperforms exis
ting person re-id methods on three occlusion datasets, while remains top perform
ance on two holistic datasets.
****

Robust Person Re-Identification by Modelling Feature Uncertainty
Tianyuan Yu, Da Li, Yongxin Yang, Timothy M. Hospedales, Tao Xiang; Proceedi
ngs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp
. 552-561
We aim to learn deep person re-identification (ReID) models that are robust agai
nst noisy training data. Two types of noise are prevalent in practice: (1) label
 noise caused by human annotator errors and (2) data outliers caused by person d
etector errors or occlusion. Both types of noise pose serious problems for train
ing ReID models, yet have been largely ignored so far. In this paper, we propose
 a novel deep network termed DistributionNet for robust ReID. Instead of represe
nting each person image as a feature vector, DistributionNet models it as a Gaus
sian distribution with its variance representing the uncertainty of the extracte
d features. A carefully designed loss is formulated in DistributionNet to uneven
ly allocate uncertainty across training samples. Consequently, noisy samples are
 assigned large variance/uncertainty, which effectively alleviates their negativ
e impacts on model fitting. Extensive experiments demonstrate that our model is
more effective than alternative noise-robust deep models. The source code is ava
ilable at: https://github.com/TianyuanYu/DistributionNet
****

Co-Segmentation Inspired Attention Networks for Video-Based Person Re-Identifica
tion
Arulkumar Subramaniam, Athira Nambiar, Anurag Mittal; Proceedings of the IEEE/
CVF International Conference on Computer Vision (ICCV), 2019, pp. 562-572
Person re-identification (Re-ID) is an important real-world surveillance problem
 that entails associating a person's identity over a network of cameras. Video-b
ased Re-ID approaches have gained significant attention recently since a video,
and not just an image, is often available. In this work, we propose a novel Co-s
egmentation inspired video Re-ID deep architecture and formulate a Co-segmentati
on based Attention Module (COSAM) that activates a common set of salient feature
s across multiple frames of a video via mutual consensus in an unsupervised mann
er. As opposed to most of the prior work, our approach is able to attend to pers
on accessories along with the person. Our plug-and-play and interpretable COSAM
module applied on two deep architectures (ResNet50, SE-ResNet50) outperform the
state-of-the-art methods on three benchmark datasets.
****

A Delay Metric for Video Object Detection: What Average Precision Fails to Tell
Huizi Mao, Xiaodong Yang, William J. Dally; Proceedings of the IEEE/CVF Intern
ational Conference on Computer Vision (ICCV), 2019, pp. 573-582
Average precision (AP) is a widely used metric to evaluate detection accuracy of
 image and video object detectors. In this paper, we analyze the object detectio
n from video and point out that mAP alone is not sufficient to capture the tempo
ral nature of video object detection. To tackle this problem, we propose a compr
ehensive metric, Average Delay (AD), to measure and compare detection delay. To
facilitate delay evaluation, we carefully select a subset of ImageNet VID, which
 we name as ImageNet VIDT with an emphasis on complex trajectories. By extensive
ly evaluating a wide range of detectors on VIDT, we show that most methods drast

ically increase the detection delay but still preserve mAP well. In other words, mAP is not sensitive enough to reflect the temporal characteristics of a video object detector. Our results suggest that video object detection methods should be evaluated with a delay metric, particularly for latency-critical applications such as autonomous vehicle perception.
****

IL2M: Class Incremental Learning With Dual Memory
Eden Belouadah, Adrian Popescu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 583-592
This paper presents a class incremental learning (IL) method which exploits fine tuning and a dual memory to reduce the negative effect of catastrophic forgetting in image recognition. First, we simplify the current fine tuning based approaches which use a combination of classification and distillation losses to compensate for the limited availability of past data. We find that the distillation term actually hurts performance when a memory is allowed. Then, we modify the usual class IL memory component. Similar to existing works, a first memory stores exemplar images of past classes. A second memory is introduced here to store past class statistics obtained when they were initially learned. The intuition here is that classes are best modeled when all their data are available and that their initial statistics are useful across different incremental states. A prediction bias towards newly learned classes appears during inference because the dataset is imbalanced in their favor. The challenge is to make predictions of new and past classes more comparable. To do this, scores of past classes are rectified by leveraging contents from both memories. The method has negligible added cost, both in terms of memory and of inference complexity. Experiments with three large public datasets show that the proposed approach is more effective than a range of competitive state-of-the-art methods.
****

Asymmetric Non-Local Neural Networks for Semantic Segmentation
Zhen Zhu, Mengde Xu, Song Bai, Tengteng Huang, Xiang Bai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 593-602
The non-local module works as a particularly useful technique for semantic segmentation while criticized for its prohibitive computation and GPU memory occupation. In this paper, we present Asymmetric Non-local Neural Network to semantic segmentation, which has two prominent components: Asymmetric Pyramid Non-local Block (APNB) and Asymmetric Fusion Non-local Block (AFNB). APNB leverages a pyramid sampling module into the non-local block to largely reduce the computation and memory consumption without sacrificing the performance. AFNB is adapted from APNB to fuse the features of different levels under a sufficient consideration of long range dependencies and thus considerably improves the performance. Extensive experiments on semantic segmentation benchmarks demonstrate the effectiveness and efficiency of our work. In particular, we report the state-of-the-art performance of 81.3 mIoU on the Cityscapes test set. For a 256x128 input, APNB is around 6 times faster than a non-local block on GPU while 28 times smaller in GPU running memory occupation. Code is available at: https://github.com/MendelXu/ANN.git.
****

CCNet: Criss-Cross Attention for Semantic Segmentation
Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, Wenyu Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 603-612
Full-image dependencies provide useful contextual information to benefit visual understanding problems. In this work, we propose a Criss-Cross Network (CCNet) for obtaining such contextual information in a more effective and efficient way. Concretely, for each pixel, a novel criss-cross attention module in CCNet harvests the contextual information of all the pixels on its criss-cross path. By taking a further recurrent operation, each pixel can finally capture the full-image dependencies from all pixels. Overall, CCNet is with the following merits: 1) GPU memory friendly. Compared with the non-local block, the proposed recurrent criss-cross attention module requires 11x less GPU memory usage. 2) High computatio

nal efficiency. The recurrent criss-cross attention significantly reduces FLOPs by about 85% of the non-local block in computing full-image dependencies. 3) The state-of-the-art performance. We conduct extensive experiments on popular semantic segmentation benchmarks including Cityscapes, ADE20K, and instance segmentation benchmark COCO. In particular, our CCNet achieves the mIoU score of 81.4 and 45.22 on Cityscapes test set and ADE20K validation set, respectively, which are the new state-of-the-art results. The source code is available at https://github.com/speedinghzl/CCNet.

****

Convex Shape Prior for Multi-Object Segmentation Using a Single Level Set Function

Shousheng Luo, Xue-Cheng Tai, Limei Huo, Yang Wang, Roland Glowinski; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 613-621

Many objects in real world have convex shapes. It is a difficult task to have representations for convex shapes with good and fast numerical solutions. This paper proposes a method to incorporate convex shape prior for multi-object segmentation using level set method. The relationship between the convexity of the segmented objects and the signed distance function corresponding to their union is analyzed theoretically. This result is combined with Gaussian mixture method for the multiple objects segmentation with convexity shape prior. Alternating direction method of multiplier (ADMM) is adopted to solve the proposed model. Special boundary conditions are also imposed to obtain efficient algorithms for 4th order partial differential equations in one step of ADMM algorithm. In addition, our method only needs one level set function regardless of the number of objects. So the increase in the number of objects does not result in the increase of model and algorithm complexity. Various numerical experiments are illustrated to show the performance and advantages of the proposed method.

****

Surface Networks via General Covers

Niv Haim, Nimrod Segol, Heli Ben-Hamu, Haggai Maron, Yaron Lipman; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 632-641

Developing deep learning techniques for geometric data is an active and fruitful research area. This paper tackles the problem of sphere-type surface learning by developing a novel surface-to-image representation. Using this representation we are able to quickly adapt successful CNN models to the surface setting. The surface-image representation is based on a covering map from the image domain to the surface. Namely, the map wraps around the surface several times, making sure that every part of the surface is well represented in the image. Differently from previous surface-to-image representations, we provide a low distortion coverage of all surface parts in a single image. Specifically, for the use case of learning spherical signals, our representation provides a low distortion alternative to several popular spherical parameterizations used in deep learning. We have used the surface-to-image representation to apply standard CNN architectures to 3D models including spherical signals. We show that our method achieves state of the art or comparable results on the tasks of shape retrieval, shape classification and semantic shape segmentation.

****

SSAP: Single-Shot Instance Segmentation With Affinity Pyramid

Naiyu Gao, Yanhu Shan, Yupei Wang, Xin Zhao, Yinan Yu, Ming Yang, Kaiqi Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 642-651

Recently, proposal-free instance segmentation has received increasing attention due to its concise and efficient pipeline. Generally, proposal-free methods generate instance-agnostic semantic segmentation labels and instance-aware features to group pixels into different object instances. However, previous methods mostly employ separate modules for these two sub-tasks and require multiple passes for inference. We argue that treating these two sub-tasks separately is suboptimal. In fact, employing multiple separate modules significantly reduces the potenti

al for application. The mutual benefits between the two complementary sub-tasks are also unexplored. To this end, this work proposes a single-shot proposal-free instance segmentation method that requires only one single pass for prediction. Our method is based on a pixel-pair affinity pyramid, which computes the probability that two pixels belong to the same instance in a hierarchical manner. The affinity pyramid can also be jointly learned with the semantic class labeling and achieve mutual benefits. Moreover, incorporating with the learned affinity pyramid, a novel cascaded graph partition module is presented to sequentially generate instances from coarse to fine. Unlike previous time-consuming graph partition methods, this module achieves 5x speedup and 9% relative improvement on Average-Precision (AP). Our approach achieves new state of the art on the challenging Cityscapes dataset.

****

Learning Propagation for Arbitrarily-Structured Data

Sifei Liu, Xueting Li, Varun Jampani, Shalini De Mello, Jan Kautz; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 652-661

Processing an input signal that contains arbitrary structures, e.g., superpixels and point clouds, remains a big challenge in computer vision. Linear diffusion, an effective model for image processing, has been recently integrated with deep learning algorithms. In this paper, we propose to learn pairwise relations among data points in a global fashion to improve semantic segmentation with arbitrarily-structured data, through spatial generalized propagation networks (SGPN). The network propagates information on a group of graphs, which represent the arbitrarily-structured data, through a learned, linear diffusion process. The module is flexible to be embedded and jointly trained with many types of networks, e.g., CNNs. We experiment with semantic segmentation networks, where we use our propagation module to jointly train on different data -- images, superpixels, and point clouds. We show that SGPN consistently improves the performance of both pixel and point cloud segmentation, compared to networks that do not contain this module. Our method suggests an effective way to model the global pairwise relations for arbitrarily-structured data.

****

MultiSeg: Semantically Meaningful, Scale-Diverse Segmentations From Minimal User Input

Jun Hao Liew, Scott Cohen, Brian Price, Long Mai, Sim-Heng Ong, Jiashi Feng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 662-670

Existing deep learning-based interactive image segmentation approaches typically assume the target-of-interest is always a single object and fail to account for the potential diversity in user expectations, thus requiring excessive user input when it comes to segmenting an object part or a group of objects instead. Motivated by the observation that the object part, full object, and a collection of objects essentially differ in size, we propose a new concept called scale-diversity, which characterizes the spectrum of segmentations w.r.t. different scales. To address this, we present MultiSeg, a scale-diverse interactive image segmentation network that incorporates a set of two-dimensional scale priors into the model to generate a set of scale-varying proposals that conform to the user input. We explicitly encourage segmentation diversity during training by synthesizing diverse training samples for a given image. As a result, our method allows the user to quickly locate the closest segmentation target for further refinement if necessary. Despite its simplicity, experimental results demonstrate that our proposed model is capable of quickly producing diverse yet plausible segmentation outputs, reducing the user interaction required, especially in cases where many types of segmentations (object parts or groups) are expected.

****

Robust Motion Segmentation From Pairwise Matches

Federica Arrigoni, Tomas Pajdla; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 671-681

In this paper we consider the problem of motion segmentation, where only pairwis

e correspondences are assumed as input without prior knowledge about tracks. The problem is formulated as a two-step process. First, motion segmentation is performed on image pairs independently. Secondly, we combine independent pairwise segmentation results in a robust way into the final globally consistent segmentation. Our approach is inspired by the success of averaging methods. We demonstrate in simulated as well as in real experiments that our method is very effective in reducing the errors in the pairwise motion segmentation and can cope with large number of mismatches.

****

InstaBoost: Boosting Instance Segmentation via Probability Map Guided Copy-Pasting

Hao-Shu Fang, Jianhua Sun, Runzhong Wang, Minghao Gou, Yong-Lu Li, Cewu Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 682-691

Instance segmentation requires a large number of training samples to achieve satisfactory performance and benefits from proper data augmentation. To enlarge the training set and increase the diversity, previous methods have investigated using data annotation from other domain (e.g. bbox, point) in a weakly supervised mechanism. In this paper, we present a simple, efficient and effective method to augment the training set using the existing instance mask annotations. Exploiting the pixel redundancy of the background, we are able to improve the performance of Mask R-CNN for 1.7 mAP on COCO dataset and 3.3 mAP on Pascal VOC dataset by simply introducing random jittering to objects. Furthermore, we propose a location probability map based approach to explore the feasible locations that objects can be placed based on local appearance similarity. With the guidance of such map, we boost the performance of R101-Mask R-CNN on instance segmentation from 35.7 mAP to 37.9 mAP without modifying the backbone or network structure. Our method is simple to implement and does not increase the computational complexity. It can be integrated into the training pipeline of any instance segmentation model without affecting the training and inference efficiency. Our code and models have been released at https://github.com/GothicAi/InstaBoost.

****

Racial Faces in the Wild: Reducing Racial Bias by Information Maximization Adaptation Network

Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, Yaohai Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 692-702

Racial bias is an important issue in biometric, but has not been thoroughly studied in deep face recognition. In this paper, we first contribute a dedicated dataset called Racial Faces in-the-Wild (RFW) database, on which we firmly validated the racial bias of four commercial APIs and four state-of-the-art (SOTA) algorithms. Then, we further present the solution using deep unsupervised domain adaptation and propose a deep information maximization adaptation network (IMAN) to alleviate this bias by using Caucasian as source domain and other races as target domains. This unsupervised method simultaneously aligns global distribution to decrease race gap at domain-level, and learns the discriminative target representations at cluster level. A novel mutual information loss is proposed to further enhance the discriminative ability of network output without label information. Extensive experiments on RFW, GBU, and IJB-A databases show that IMAN successfully learns features that generalize well across different races and across different databases.

****

Uncertainty Modeling of Contextual-Connections Between Tracklets for Unconstrained Video-Based Face Recognition

Jingxiao Zheng, Ruichi Yu, Jun-Cheng Chen, Boyu Lu, Carlos D. Castillo, Rama Chellappa; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 703-712

Unconstrained video-based face recognition is a challenging problem due to significant within-video variations caused by pose, occlusion and blur. To tackle this problem, an effective idea is to propagate the identity from high-quality face

s to low-quality ones through contextual connections, which are constructed base d on context such as body appearance. However, previous methods have often propa gated erroneous information due to lack of uncertainty modeling of the noisy con textual connections. In this paper, we propose the Uncertainty-Gated Graph (UGG) , which conducts graph-based identity propagation between tracklets, which are r epresented by nodes in a graph. UGG explicitly models the uncertainty of the con textual connections by adaptively updating the weights of the edge gates accordi ng to the identity distributions of the nodes during inference. UGG is a generic graphical model that can be applied at only inference time or with end-to-end t raining. We demonstrate the effectiveness of UGG with state-of-the-art results i n the recently released challenging Cast Search in Movies and IARPA Janus Survei llance Video Benchmark dataset.

****

Spatio-Temporal Fusion Based Convolutional Sequence Learning for Lip Reading

Xingxuan Zhang, Feng Cheng, Shilin Wang; Proceedings of the IEEE/CVF Internati onal Conference on Computer Vision (ICCV), 2019, pp. 713-722

Current state-of-the-art approaches for lip reading are based on sequence-to-seq uence architectures that are designed for natural machine translation and audio speech recognition. Hence, these methods do not fully exploit the characteristic s of the lip dynamics, causing two main drawbacks. First, the short-range tempor al dependencies, which are critical to the mapping from lip images to visemes, r eceives no extra attention. Second, local spatial information is discarded in th e existing sequence models due to the use of global average pooling (GAP). To we ll solve these drawbacks, we propose a Temporal Focal block to sufficiently desc ribe short-range dependencies and a Spatio-Temporal Fusion Module (STFM) to main tain the local spatial information and to reduce the feature dimensions as well. From the experiment results, it is demonstrated that our method achieves compar able performance with the state-of-the-art approach using much less training dat a and much lighter Convolutional Feature Extractor. The training time is reduced by 12 days due to the convolutional structure and the local self-attention mech anism.

****

Occlusion-Aware Networks for 3D Human Pose Estimation in Video

Yu Cheng, Bo Yang, Bo Wang, Wending Yan, Robby T. Tan; Proceedings of the IE EE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 723-732

Occlusion is a key problem in 3D human pose estimation from a monocular video. T o address this problem, we introduce an occlusion-aware deep-learning framework. By employing estimated 2D confidence heatmaps of keypoints and an optical-flow consistency constraint, we filter out the unreliable estimations of occluded key points. When occlusion occurs, we have incomplete 2D keypoints and feed them to our 2D and 3D temporal convolutional networks (2D and 3D TCNs) that enforce temp oral smoothness to produce a complete 3D pose. By using incomplete 2D keypoints, instead of complete but incorrect ones, our networks are less affected by the e rror-prone estimations of occluded keypoints. Training the occlusion-aware 3D TC N requires pairs of a 3D pose and a 2D pose with occlusion labels. As no such a dataset is available, we introduce a "Cylinder Man Model" to approximate the occ upation of body parts in 3D space. By projecting the model onto a 2D plane in di fferent viewing angles, we obtain and label the occluded keypoints, providing us plenty of training data. In addition, we use this model to create a pose regula rization constraint, preferring the 2D estimations of unreliable keypoints to be occluded. Our method outperforms state-of-the-art methods on Human 3.6M and Hum anEva-I datasets.

****

Context-Aware Feature and Label Fusion for Facial Action Unit Intensity Estimati on With Partially Labeled Data

Yong Zhang, Haiyong Jiang, Baoyuan Wu, Yanbo Fan, Qiang Ji; Proceedings of t he IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 733-74 2

Facial action unit (AU) intensity estimation is a fundamental task for facial be haviour analysis. Most previous methods use a whole face image as input for inte

nsity prediction. Considering that AUs are defined according to their correspond ing local appearance, a few patch-based methods utilize image features of local patches. However, fusion of local features is always performed via straightforwa rd feature concatenation or summation. Besides, these methods require fully anno tated databases for model learning, which is expensive to acquire. In this paper , we propose a novel weakly supervised patch-based deep model on basis of two ty pes of attention mechanisms for joint intensity estimation of multiple AUs. The model consists of a feature fusion module and a label fusion module. And we augm ent attention mechanisms of these two modules with a learnable task-related cont ext, as one patch may play different roles in analyzing different AUs and each A U has its own temporal evolution rule. The context-aware feature fusion module i s used to capture spatial relationships among local patches while the context-aw are label fusion module is used to capture the temporal dynamics of AUs. The lat ter enables the model to be trained on a partially annotated database. Experimen tal evaluations on two benchmark expression databases demonstrate the superior p erformance of the proposed method.

****

Distill Knowledge From NRSfM for Weakly Supervised 3D Pose Learning

Chaoyang Wang, Chen Kong, Simon Lucey; Proceedings of the IEEE/CVF Internation al Conference on Computer Vision (ICCV), 2019, pp. 743-752

We propose to learn a 3D pose estimator by distilling knowledge from Non-Rigid S tructure from Motion (NRSfM). Our method uses solely 2D landmark annotations. No 3D data, multi-view/temporal footage, or object specific prior is required. Thi s alleviates the data bottleneck, which is one of the major concern for supervis ed methods. The challenge for using NRSfM as teacher is that they often make poo r depth reconstruction when the 2D projections have strong ambiguity. Directly u sing those wrong depth as hard target would negatively impact the student. Inste ad, we propose a novel loss that ties depth prediction to the cost function used in NRSfM. This gives the student pose estimator freedom to reduce depth error b y associating with image features. Validated on H3.6M dataset, our learned 3D po se estimation network achieves more accurate reconstruction compared to NRSfM me thods. It also outperforms other weakly supervised methods, in spite of using si gnificantly less supervision.

****

MONET: Multiview Semi-Supervised Keypoint Detection via Epipolar Divergence

Yuan Yao, Yasamin Jafarian, Hyun Soo Park; Proceedings of the IEEE/CVF Interna tional Conference on Computer Vision (ICCV), 2019, pp. 753-762

This paper presents MONET---an end-to-end semi-supervised learning framework for a keypoint detector using multiview image streams. In particular, we consider g eneral subjects such as non-human species where attaining a large scale annotate d dataset is challenging. While multiview geometry can be used to self-supervise the unlabeled data, integrating the geometry into learning a keypoint detector is challenging due to representation mismatch. We address this mismatch by formu lating a new differentiable representation of the epipolar constraint called epi polar divergence---a generalized distance from the epipolar lines to the corresp onding keypoint distribution. Epipolar divergence characterizes when two view ke ypoint distributions produce zero reprojection error. We design a twin network t hat minimizes the epipolar divergence through stereo rectification that can sign ificantly alleviate computational complexity and sampling aliasing in training. We demonstrate that our framework can localize customized keypoints of diverse s pecies, e.g., humans, dogs, and monkeys.

****

Talking With Hands 16.2M: A Large-Scale Dataset of Synchronized Body-Finger Moti on and Audio for Conversational Motion Analysis and Synthesis

Gilwoo Lee, Zhiwei Deng, Shugao Ma, Takaaki Shiratori, Siddhartha S. Srinива sa, Yaser Sheikh; Proceedings of the IEEE/CVF International Conference on Compu ter Vision (ICCV), 2019, pp. 763-772

We present a 16.2-million frame (50-hour) multimodal dataset of two-person face- to-face spontaneous conversations. Our dataset features synchronized body and fi nger motion as well as audio data. To the best of our knowledge, it represents t

he largest motion capture and audio dataset of natural conversations to date. Th e statistical analysis verifies strong intraperson and interperson covariance of arm, hand, and speech features, potentially enabling new directions on data-dri ven social behavior analysis, prediction, and synthesis. As an illustration, we propose a novel real-time finger motion synthesis method: a temporal neural netw ork innovatively trained with an inverse kinematics (IK) loss, which adds skelet al structural information to the generative model. Our qualitative user study sh ows that the finger motion generated by our method is perceived as natural and c onversation enhancing, while the quantitative ablation study demonstrates the ef fectiveness of IK loss.

****

Occlusion Robust Face Recognition Based on Mask Learning With Pairwise Different ial Siamese Network

Lingxue Song, Dihong Gong, Zhifeng Li, Changsong Liu, Wei Liu; Proceedings o f the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 773 -782

Deep Convolutional Neural Networks (CNNs) have been pushing the frontier of face recognition over past years. However, existing CNN models are far less accurate when handling partially occluded faces. These general face models generalize po orly for occlusions on variable facial areas. Inspired by the fact that human vi sual system explicitly ignores the occlusion and only focuses on the non-occlude d facial areas, we propose a mask learning strategy to find and discard corrupte d feature elements from recognition. A mask dictionary is firstly established by exploiting the differences between the top conv features of occluded and occlus ion-free face pairs using innovatively designed pairwise differential siamese ne twork (PDSN). Each item of this dictionary captures the correspondence between o ccluded facial areas and corrupted feature elements, which is named Feature Disc arding Mask (FDM). When dealing with a face image with random partial occlusions , we generate its FDM by combining relevant dictionary items and then multiply i t with the original features to eliminate those corrupted feature elements from recognition. Comprehensive experiments on both synthesized and realistic occlude d face datasets show that the proposed algorithm significantly outperforms the s tate-of-the-art systems.

****

Teacher Supervises Students How to Learn From Partially Labeled Images for Facia l Landmark Detection

Xuanyi Dong, Yi Yang; Proceedings of the IEEE/CVF International Conference on C omputer Vision (ICCV), 2019, pp. 783-792

Facial landmark detection aims to localize the anatomically defined points of hu man faces. In this paper, we study facial landmark detection from partially labe led facial images. A typical approach is to (1) train a detector on the labeled images; (2) generate new training samples using this detector's prediction as ps eudo labels of unlabeled images; (3) retrain the detector on the labeled samples and partial pseudo labeled samples. In this way, the detector can learn from bo th labeled and unlabeled data and become robust. In this paper, we propose an in teraction mechanism between a teacher and two students to generate more reliable pseudo labels for unlabeled data, which are beneficial to semi-supervised facia l landmark detection. Specifically, the two students are instantiated as dual de tectors. The teacher learns to judge the quality of the pseudo labels generated by the students and filter out unqualified samples before the retraining stage. In this way, the student detectors get feedback from their teacher and are retra ined by premium data generated by itself. Since the two students are trained by different samples, a combination of their predictions will be more robust as the final prediction compared to either prediction. Extensive experiments on 300-W and AFLW benchmarks show that the interactions between teacher and students cont ribute to better utilization of the unlabeled data and achieves state-of-the-art performance.

****

A2J: Anchor-to-Joint Regression Network for 3D Articulated Pose Estimation From a Single Depth Image

Fu Xiong, Boshen Zhang, Yang Xiao, Zhiguo Cao, Taidong Yu, Joey Tianyi Zhou , Junsong Yuan; Proceedings of the IEEE/CVF International Conference on Compute r Vision (ICCV), 2019, pp. 793-802

For 3D hand and body pose estimation task in depth image, a novel anchor-based a pproach termed Anchor-to-Joint regression network (A2J) with the end-to-end lear ning ability is proposed. Within A2J, anchor points able to capture global-local spatial context information are densely set on depth image as local regressors for the joints. They contribute to predict the positions of the joints in ensemb le way to enhance generalization ability. The proposed 3D articulated pose estim ation paradigm is different from the state-of-the-art encoder-decoder based FCN, 3D CNN and point-set based manners. To discover informative anchor points towar ds certain joint, anchor proposal procedure is also proposed for A2J. Meanwhile 2D CNN (i.e., ResNet- 50) is used as backbone network to drive A2J, without usin g time-consuming 3D convolutional or deconvolutional layers. The experiments on 3 hand datasets and 2 body datasets verify A2J's superiority. Meanwhile, A2J is of high running speed around 100 FPS on single NVIDIA 1080Ti GPU.

****

TexturePose: Supervising Human Mesh Estimation With Texture Consistency

Georgios Pavlakos, Nikos Kolotouros, Kostas Daniilidis; Proceedings of the IEE E/CVF International Conference on Computer Vision (ICCV), 2019, pp. 803-812

This work addresses the problem of model-based human pose estimation. Recent app roaches have made significant progress towards regressing the parameters of para metric human body models directly from images. Because of the absence of images with 3D shape ground truth, relevant approaches rely on 2D annotations or sophis ticated architecture designs. In this work, we advocate that there are more cues we can leverage, which are available for free in natural images, i.e., without getting more annotations, or modifying the network architecture. We propose a na tural form of supervision, that capitalizes on the appearance constancy of a per son among different frames (or viewpoints). This seemingly insignificant and oft en overlooked cue goes a long way for model-based pose estimation. The parametri c model we employ allows us to compute a texture map for each frame. Assuming th at the texture of the person does not change dramatically between frames, we can apply a novel texture consistency loss, which enforces that each point in the t exture map has the same texture value across all frames. Since the texture is tr ansferred in this common texture map space, no camera motion computation is nece ssary, or even an assumption of smoothness among frames. This makes our proposed supervision applicable in a variety of settings, ranging from monocular video, to multi-view images. We benchmark our approach against strong baselines that re quire the same or even more annotations that we do and we consistently outperfor m them. Simultaneously, we achieve state-of-the-art results among model-based po se estimation approaches in different benchmarks. The project website with video s, results, and code can be found at https://seas.upenn.edu/ pavlakos/projects/t exturepose.

****

FreiHAND: A Dataset for Markerless Capture of Hand Pose and Shape From Single RG B Images

Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, T homas Brox; Proceedings of the IEEE/CVF International Conference on Computer Vis ion (ICCV), 2019, pp. 813-822

Estimating 3D hand pose from single RGB images is a highly ambiguous problem tha t relies on an unbiased training dataset. In this paper, we analyze cross-datase t generalization when training on existing datasets. We find that approaches per form well on the datasets they are trained on, but do not generalize to other da tasets or in-the-wild scenarios. As a consequence, we introduce the first large- scale, multi-view hand dataset that is accompanied by both 3D hand pose and shap e annotations. For annotating this real-world dataset, we propose an iterative, semi-automated `human-in-the-loop' approach, which includes hand fitting optimiz ation to infer both the 3D pose and shape for each sample. We show that methods trained on our dataset consistently perform well when tested on other datasets. Moreover, the dataset allows us to train a network that predicts the full articu

lated hand shape from a single RGB image. The evaluation set can serve as a benc
hmark for articulated hand shape estimation.
****

Markerless Outdoor Human Motion Capture Using Multiple Autonomous Micro Aerial V
ehicles
Nitin Saini, Eric Price, Rahul Tallamraju, Raffi Enficiaud, Roman Ludwig, I
gor Martinovic, Aamir Ahmad, Michael J. Black; Proceedings of the IEEE/CVF Int
ernational Conference on Computer Vision (ICCV), 2019, pp. 823-832
Capturing human motion in natural scenarios means moving motion capture out of t
he lab and into the wild. Typical approaches rely on fixed, calibrated, cameras
and reflective markers on the body, significantly limiting the motions that can
be captured. To make motion capture truly unconstrained, we describe the first f
ully autonomous outdoor capture system based on flying vehicles. We use multiple
micro-aerial-vehicles(MAVs), each equipped with a monocular RGB camera, an IMU,
and a GPS receiver module. These detect the person, optimize their position, an
d localize themselves approximately. We then develop a markerless motion capture
method that is suitable for this challenging scenario with a distant subject, v
iewed from above, with approximately calibrated and moving cameras. We combine m
ultiple state-of-the-art 2D joint detectors with a 3D human body model and a pow
erful prior on human pose. We jointly optimize for 3D body pose and camera pose
to robustly fit the 2D measurements. To our knowledge, this is the first success
ful demonstration of outdoor, full-body, markerless motion capture from autonomo
us flying vehicles.
****

Toyota Smarthome: Real-World Activities of Daily Living
Srijan Das, Rui Dai, Michal Koperski, Luca Minciullo, Lorenzo Garattoni, Fr
ancois Bremond, Gianpiero Francesca; Proceedings of the IEEE/CVF International
Conference on Computer Vision (ICCV), 2019, pp. 833-842
The performance of deep neural networks is strongly influenced by the quantity a
nd quality of annotated data. Most of the large activity recognition datasets co
nsist of data sourced from the web, which does not reflect challenges that exist
in activities of daily living. In this paper, we introduce a large real-world v
ideo dataset for activities of daily living: Toyota Smarthome. The dataset consi
sts of 16K RGB+D clips of 31 activity classes, performed by seniors in a smartho
me. Unlike previous datasets, videos were fully unscripted. As a result, the dat
aset poses several challenges: high intra-class variation, high class imbalance,
simple and composite activities, and activities with similar motion and variabl
e duration. Activities were annotated with both coarse and fine-grained labels.
These characteristics differentiate Toyota Smarthome from other datasets for act
ivity recognition. As recent activity recognition approaches fail to address the
challenges posed by Toyota Smarthome, we present a novel activity recognition m
ethod with attention mechanism. We propose a pose driven spatio-temporal attenti
on mechanism through 3D ConvNets. We show that our novel method outperforms stat
e-of-the-art methods on benchmark datasets, as well as on the Toyota Smarthome d
ataset. We release the dataset for research use.
****

Relation Parsing Neural Network for Human-Object Interaction Detection
Penghao Zhou, Mingmin Chi; Proceedings of the IEEE/CVF International Conference
on Computer Vision (ICCV), 2019, pp. 843-851
Human-Object Interaction Detection devotes to infer a triplet < human, verb, obj
ect > between human and objects. In this paper, we propose a novel model, i.e.,
Relation Parsing Neural Network (RPNN), to detect human-object interactions. Spe
cifically, the network is represented by two graphs, i.e., Object-Bodypart Graph
and Human-Bodypart Graph. Here, the Object-Bodypart Graph dynamically captures
the relationship between body parts and the surrounding objects. The Human-Bodyp
art Graph infers the relationship between human and body parts, and assembles bo
dy part contexts to predict actions. These two graphs are associated through an
action passing mechanism. The proposed RPNN model is able to implicitly parse a
pairwise relation in two graphs without supervised labels. Experiments conducted
on V-COCO and HICO-DET datasets confirm the effectiveness of the proposed RPNN

network which significantly outperforms state-of-the-art methods.
****

DistInit: Learning Video Representations Without a Single Labeled Video
Rohit Girdhar, Du Tran, Lorenzo Torresani, Deva Ramanan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 852-861
Video recognition models have progressed significantly over the past few years, evolving from shallow classifiers trained on hand-crafted features to deep spatiotemporal networks. However, labeled video data required to train such models has not been able to keep up with the ever increasing depth and sophistication of these networks. In this work we propose an alternative approach to learning video representations that requires no semantically labeled videos, and instead leverages the years of effort in collecting and labeling large and clean still-image datasets. We do so by using state-of-the-art models pre-trained on image datasets as "teachers" to train video models in a distillation framework. We demonstrate that our method learns truly spatiotemporal features, despite being trained only using supervision from still-image networks. Moreover, it learns good representations across different input modalities, using completely uncurated raw video data sources and with different 2D teacher models. Our method obtains strong transfer performance, outperforming standard techniques for bootstrapping video architectures with image based models by 16%. We believe that our approach opens up new approaches for learning spatiotemporal representations from unlabeled video data.
****

Zero-Shot Anticipation for Instructional Activities
Fadime Sener, Angela Yao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 862-871
How can we teach a robot to predict what will happen next for an activity it has never seen before? We address the problem of zero-shot anticipation by presenting a hierarchical model that generalizes instructional knowledge from large-scale text-corpora and transfers the knowledge to the visual domain. Given a portion of an instructional video, our model predicts coherent and plausible actions multiple steps into the future, all in rich natural language. To demonstrate the anticipation capabilities of our model, we introduce the Tasty Videos dataset, a collection of 2511 recipes for zero-shot learning, recognition and anticipation.
****

Making the Invisible Visible: Action Recognition Through Walls and Occlusions
Tianhong Li, Lijie Fan, Mingmin Zhao, Yingcheng Liu, Dina Katabi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 872-881
Understanding people's actions and interactions typically depends on seeing them. Automating the process of action recognition from visual data has been the topic of much research in the computer vision community. But what if it is too dark, or if the person is occluded or behind a wall? In this paper, we introduce a neural network model that can detect human actions through walls and occlusions, and in poor lighting conditions. Our model takes radio frequency (RF) signals as input, generates 3D human skeletons as an intermediate representation, and recognizes actions and interactions of multiple people over time. By translating the input to an intermediate skeleton-based representation, our model can learn from both vision-based and RF-based datasets, and allow the two tasks to help each other. We show that our model achieves comparable accuracy to vision-based action recognition systems in visible scenarios, yet continues to work accurately when people are not visible, hence addressing scenarios that are beyond the limit of today's vision-based action recognition.
****

Recursive Visual Sound Separation Using Minus-Plus Net
Xudong Xu, Bo Dai, Dahua Lin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 882-891
Sounds provide rich semantics, complementary to visual data, for many tasks. However, in practice, sounds from multiple sources are often mixed together. In this paper we propose a novel framework, referred to as MinusPlus Network (MP-Net),

for the task of visual sound separation. MP-Net separates sounds recursively in the order of average energy, removing the separated sound from the mixture at the end of each prediction, until the mixture becomes empty or contains only noise. In this way, MP-Net could be applied to sound mixtures with arbitrary numbers and types of sounds. Moreover, while MP-Net keeps removing sounds with large energy from the mixture, sounds with small energy could emerge and become clearer, so that the separation is more accurate. Compared to previous methods, MP-Net obtains state-of-the-art results on two large scale datasets, across mixtures with different types and numbers of sounds.
****

Unsupervised Video Interpolation Using Cycle Consistency
Fitsum A. Reda, Deqing Sun, Aysegul Dundar, Mohammad Shoeybi, Guilin Liu, Kevin J. Shih, Andrew Tao, Jan Kautz, Bryan Catanzaro; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 892-900
Learning to synthesize high frame rate videos via interpolation requires large quantities of high frame rate training videos, which, however, are scarce, especially at high resolutions. Here, we propose unsupervised techniques to synthesize high frame rate videos directly from low frame rate videos using cycle consistency. For a triplet of consecutive frames, we optimize models to minimize the discrepancy between the center frame and its cycle reconstruction, obtained by interpolating back from interpolated intermediate frames. This simple unsupervised constraint alone achieves results comparable with supervision using the ground truth intermediate frames. We further introduce a pseudo supervised loss term that enforces the interpolated frames to be consistent with predictions of a pre-trained interpolation model. The pseudo supervised loss term, used together with cycle consistency, can effectively adapt a pre-trained model to a new target domain. With no additional data and in a completely unsupervised fashion, our techniques significantly improve pre-trained models on new target domains, increasing PSNR values from 32.84dB to 33.05dB on the Slowflow and from 31.82dB to 32.53dB on the Sintel evaluation datasets.
****

Deformable Surface Tracking by Graph Matching
Tao Wang, Haibin Ling, Congyan Lang, Songhe Feng, Xiaohui Hou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 901-910
This paper addresses the problem of deformable surface tracking from monocular images. Specifically, we propose a graph-based approach that effectively explores the structure information of the surface to enhance tracking performance. Our approach solves simultaneously for feature correspondence, outlier rejection and shape reconstruction by optimizing a single objective function, which is defined by means of pairwise projection errors between graph structures instead of unary projection errors between matched points. Furthermore, an efficient matching algorithm is developed based on soft matching relaxation. For evaluation, our approach is extensively compared to state-of-the-art algorithms on a standard dataset of occluded surfaces, as well as a newly compiled dataset of different surfaces with rich, weak or repetitive texture. Experimental results reveal that our approach achieves robust tracking results for surfaces with different types of texture, and outperforms other algorithms in both accuracy and efficiency.
****

Deep Meta Learning for Real-Time Target-Aware Visual Tracking
Janghoon Choi, Junseok Kwon, Kyoung Mu Lee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 911-920
In this paper, we propose a novel on-line visual tracking framework based on the Siamese matching network and meta-learner network, which run at real-time speeds. Conventional deep convolutional feature-based discriminative visual tracking algorithms require continuous re-training of classifiers or correlation filters, which involve solving complex optimization tasks to adapt to the new appearance of a target object. To alleviate this complex process, our proposed algorithm incorporates and utilizes a meta-learner network to provide the matching network with new appearance information of the target objects by adding target-aware fea

ture space. The parameters for the target-specific feature space are provided in stantly from a single forward-pass of the meta-learner network. By eliminating t he necessity of continuously solving complex optimization tasks in the course of tracking, experimental results demonstrate that our algorithm performs at a rea l-time speed while maintaining competitive performance among other state-of-the- art tracking algorithms.
****

Looking to Relations for Future Trajectory Forecast
Chiho Choi, Behzad Dariush; Proceedings of the IEEE/CVF International Conferenc e on Computer Vision (ICCV), 2019, pp. 921-930
Inferring relational behavior between road users as well as road users and their surrounding physical space is an important step toward effective modeling and p rediction of navigation strategies adopted by participants in road scenes. To th is end, we propose a relation-aware framework for future trajectory forecast. Ou r system aims to infer relational information from the interactions of road user s with each other and with the environment. The first module involves visual enc oding of spatio-temporal features, which captures human-human and human-space in teractions over time. The following module explicitly constructs pair-wise relat ions from spatio-temporal interactions and identifies more descriptive relations that highly influence future motion of the target road user by considering its past trajectory. The resulting relational features are used to forecast future l ocations of the target, in the form of heatmaps with an additional guidance of s patial dependencies and consideration of the uncertainty. Extensive evaluations on the public benchmark datasets demonstrate the robustness and efficacy of the proposed framework as observed by performances higher than the state-of-the-art methods.
****

Anchor Diffusion for Unsupervised Video Object Segmentation
Zhao Yang, Qiang Wang, Luca Bertinetto, Weiming Hu, Song Bai, Philip H. S. Torr; Proceedings of the IEEE/CVF International Conference on Computer Vision (I CCV), 2019, pp. 931-940
Unsupervised video object segmentation has often been tackled by methods based o n recurrent neural networks and optical flow. Despite their complexity, these ki nds of approach tend to favour short-term temporal dependencies and are thus pro ne to accumulating inaccuracies, which cause drift over time. Moreover, simple (static) image segmentation models, alone, can perform competitively against thes e methods, which further suggests that the way temporal dependencies are modelle d should be reconsidered. Motivated by these observations, in this paper we expl ore simple yet effective strategies to model long-term temporal dependencies. In spired by the non-local operators, we introduce a technique to establish dense c orrespondences between pixel embeddings of a reference "anchor" frame and the cu rrent one. This allows the learning of pairwise dependencies at arbitrarily long distances without conditioning on intermediate frames. Without online supervisi on, our approach can suppress the background and precisely segment the foregroun d object even in challenging scenarios, while maintaining consistent performance over time. With a mean IoU of 81.7%, our method ranks first on the DAVIS-2016 l eaderboard of unsupervised methods, while still being competitive against state- of-the-art online semi-supervised approaches. We further evaluate our method on the FBMS dataset and the video saliency dataset ViSal, showing results competiti ve with the state of the art.
****

Tracking Without Bells and Whistles
Philipp Bergmann, Tim Meinhardt, Laura Leal-Taixe; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 941-951
The problem of tracking multiple objects in a video sequence poses several chall enging tasks. For tracking-by-detection, these include object re-identification, motion prediction and dealing with occlusions. We present a tracker (without be lls and whistles) that accomplishes tracking without specifically targeting any of these tasks, in particular, we perform no training or optimization on trackin g data. To this end, we exploit the bounding box regression of an object detecto

r to predict the position of an object in the next frame, thereby converting a d etector into a Tracktor. We demonstrate the potential of Tracktor and provide a new state-of-the-art on three multi-object tracking benchmarks by extending it w ith a straightforward re-identification and camera motion compensation. We then perform an analysis on the performance and failure cases of several state-of-the -art tracking methods in comparison to our Tracktor. Surprisingly, none of the d edicated tracking methods are considerably better in dealing with complex tracki ng scenarios, namely, small and occluded objects or missing detections. However, our approach tackles most of the easy tracking scenarios. Therefore, we motivat e our approach as a new tracking paradigm and point out promising future researc h directions. Overall, Tracktor yields superior tracking performance than any cu rrent tracking method and our analysis exposes remaining and unsolved tracking c hallenges to inspire future research directions.
****

Perspective-Guided Convolution Networks for Crowd Counting
Zhaoyi Yan, Yuchen Yuan, Wangmeng Zuo, Xiao Tan, Yezhen Wang, Shilei Wen, Errui Ding; Proceedings of the IEEE/CVF International Conference on Computer Vis ion (ICCV), 2019, pp. 952-961
In this paper, we propose a novel perspective-guided convolution (PGC) for convo lutional neural network (CNN) based crowd counting (i.e. PGCNet), which aims to overcome the dramatic intra-scene scale variations of people due to the perspect ive effect. While most state-of-the-arts adopt multi-scale or multi-column archi tectures to address such issue, they generally fail in modeling continuous scale variations since only discrete representative scales are considered. PGCNet, on the other hand, utilizes perspective information to guide the spatially variant smoothing of feature maps before feeding them to the successive convolutions. A n effective perspective estimation branch is also introduced to PGCNet, which ca n be trained in either supervised setting or weakly-supervised setting when the branch has been pre-trained. Our PGCNet is single-column with moderate increase in computation, and extensive experimental results on four benchmark datasets sh ow the improvements of our method against the state-of-the-arts. Additionally, w e also introduce Crowd Surveillance, a large scale dataset for crowd counting th at contains 13,000+ high-resolution images with challenging scenarios. Code is a vailable at https://github.com/Zhaoyi-Yan/PGCNet.
****

End-to-End Wireframe Parsing
Yichao Zhou, Haozhi Qi, Yi Ma; Proceedings of the IEEE/CVF International Confe rence on Computer Vision (ICCV), 2019, pp. 962-971
We present a conceptually simple yet effective algorithm to detect wireframes in a given image. Compared to the previous methods which first predict an intermed iate heat map and then extract straight lines with heuristic algorithms, our met hod is end-to-end trainable and can directly output a vectorized wireframe that contains semantically meaningful and geometrically salient junctions and lines. To better understand the quality of the outputs, we propose a new metric for wir eframe evaluation that penalizes overlapped line segments and incorrect line con nectivities. We conduct extensive experiments and show that our method significa ntly outperforms the previous state-of-the-art wireframe and line extraction alg orithms. We hope our simple approach can be served as a baseline for future wire frame parsing studies. Code has been made publicly available at https://github.c om/zhou13/lcnn.
****

Incremental Class Discovery for Semantic Segmentation With RGBD Sensing
Yoshikatsu Nakajima, Byeongkeun Kang, Hideo Saito, Kris Kitani; Proceedings o f the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 972 -981
This work addresses the task of open world semantic segmentation using RGBD sens ing to discover new semantic classes over time. Although there are many types of objects in the real-word, current semantic segmentation methods make a closed w orld assumption and are trained only to segment a limited number of object class es. Towards a more open world approach, we propose a novel method that increment

ally learns new classes for image segmentation. The proposed system first segmen ts each RGBD frame using both color and geometric information, and then aggregat es that information to build a single segmented dense 3D map of the environment. The segmented 3D map representation is a key component of our approach as it is used to discover new object classes by identifying coherent regions in the 3D m ap that have no semantic label. The use of coherent region in the 3D map as a pr imitive element, rather than traditional elements such as surfels or voxels, als o significantly reduces the computational complexity and memory use of our metho d. It thus leads to semi-real-time performance at 10.7 Hz when incrementally upd ating the dense 3D map at every frame. Through experiments on the NYUDv2 dataset , we demonstrate that the proposed method is able to correctly cluster objects o f both known and unseen classes. We also show the quantitative comparison with t he state-of-the-art supervised methods, the processing time of each step, and th e influences of each component.

****

SSF-DAN: Separated Semantic Feature Based Domain Adaptation Network for Semantic Segmentation

Liang Du, Jingang Tan, Hongye Yang, Jianfeng Feng, Xiangyang Xue, Qibao Zhe ng, Xiaoqing Ye, Xiaolin Zhang; Proceedings of the IEEE/CVF International Conf erence on Computer Vision (ICCV), 2019, pp. 982-991

Despite the great success achieved by supervised fully convolutional models in s emantic segmentation, training the models requires a large amount of labor-inten sive work to generate pixel-level annotations. Recent works exploit synthetic da ta to train the model for semantic segmentation, but the domain adaptation betwe en real and synthetic images remains a challenging problem. In this work, we pro pose a Separated Semantic Feature based domain adaptation network, named SSF-DAN , for semantic segmentation. First, a Semantic-wise Separable Discriminator (SS- D) is designed to independently adapt semantic features across the target and so urce domains, which addresses the inconsistent adaptation issue in the class-wis e adversarial learning. In SS-D, a progressive confidence strategy is included t o achieve a more reliable separation. Then, an efficient Class-wise Adversarial loss Reweighting module (CA-R) is introduced to balance the class-wise adversari al learning process, which leads the generator to focus more on poorly adapted c lasses. The presented framework demonstrates robust performance, superior to sta te-of-the-art methods on benchmark datasets.

****

SpaceNet MVOI: A Multi-View Overhead Imagery Dataset

Nicholas Weir, David Lindenbaum, Alexei Bastidas, Adam Van Etten, Sean McPhe rson, Jacob Shermeyer, Varun Kumar, Hanlin Tang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 992-1001

Detection and segmentation of objects in overheard imagery is a challenging task . The variable density, random orientation, small size, and instance-to-instance heterogeneity of objects in overhead imagery calls for approaches distinct from existing models designed for natural scene datasets. Though new overhead imager y datasets are being developed, they almost universally comprise a single view t aken from directly overhead ("at nadir"), failing to address a critical variable : look angle. By contrast, views vary in real-world overhead imagery, particular ly in dynamic scenarios such as natural disasters where first looks are often ov er 40 degrees off-nadir. This represents an important challenge to computer visi on methods, as changing view angle adds distortions, alters resolution, and chan ges lighting. At present, the impact of these perturbations for algorithmic dete ction and segmentation of objects is untested. To address this problem, we prese nt an open source Multi-View Overhead Imagery dataset, termed SpaceNet MVOI, wit h 27 unique looks from a broad range of viewing angles (-32.5 degrees to 54.0 de grees). Each of these images cover the same 665 square km geographic extent and are annotated with 126,747 building footprint labels, enabling direct assessment of the impact of viewpoint perturbation on model performance. We benchmark mult iple leading segmentation and object detection models on: (1) building detection , (2) generalization to unseen viewing angles and resolutions, and (3) sensitivi ty of building footprint extraction to changes in resolution. We find that state

of the art segmentation and object detection models struggle to identify buildings in off-nadir imagery and generalize poorly to unseen views, presenting an important benchmark to explore the broadly relevant challenge of detecting small, heterogeneous target objects in visually dynamic contexts.
****

Multi-Level Bottom-Top and Top-Bottom Feature Fusion for Crowd Counting
Vishwanath A. Sindagi, Vishal M. Patel; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1002-1012
Crowd counting presents enormous challenges in the form of large variation in scales within images and across the dataset. These issues are further exacerbated in highly congested scenes. Approaches based on straightforward fusion of multi-scale features from a deep network seem to be obvious solutions to this problem. However, these fusion approaches do not yield significant improvements in the case of crowd counting in congested scenes. This is usually due to their limited abilities in effectively combining the multi-scale features for problems like crowd counting. To overcome this, we focus on how to efficiently leverage information present in different layers of the network. Specifically, we present a network that involves: (i) a multi-level bottom-top and top-bottom fusion (MBTTBF) method to combine information from shallower to deeper layers and vice versa at multiple levels, (ii) scale complementary feature extraction blocks (SCFB) involving cross-scale residual functions to explicitly enable flow of complementary features from adjacent conv layers along the fusion paths. Furthermore, in order to increase the effectiveness of the multi-scale fusion, we employ a principled way of generating scale-aware ground-truth density maps for training. Experiments conducted on three datasets that contain highly congested scenes (ShanghaiTech, UCF_CC_50, and UCF-QNRF) demonstrate that the proposed method is able to outperform several recent methods in all the datasets
****

Learning Lightweight Lane Detection CNNs by Self Attention Distillation
Yuenan Hou, Zheng Ma, Chunxiao Liu, Chen Change Loy; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1013-1021
Training deep models for lane detection is challenging due to the very subtle and sparse supervisory signals inherent in lane annotations. Without learning from much richer context, these models often fail in challenging scenarios, e.g., severe occlusion, ambiguous lanes, and poor lighting conditions. In this paper, we present a novel knowledge distillation approach, i.e., Self Attention Distillation (SAD), which allows a model to learn from itself and gains substantial improvement without any additional supervision or labels. Specifically, we observe that attention maps extracted from a model trained to a reasonable level would encode rich contextual information. The valuable contextual information can be used as a form of 'free' supervision for further representation learning through performing top- down and layer-wise attention distillation within the net- work itself. SAD can be easily incorporated in any feed- forward convolutional neural networks (CNN) and does not increase the inference time. We validate SAD on three popular lane detection benchmarks (TuSimple, CULane and BDD100K) using lightweight models such as ENet, ResNet- 18 and ResNet-34. The lightest model, ENet-SAD, performs comparatively or even surpasses existing algorithms. Notably, ENet-SAD has 20 x fewer parameters and runs 10 x faster compared to the state-of-the-art SCNN, while still achieving compelling performance in all benchmarks.
****

SplitNet: Sim2Sim and Task2Task Transfer for Embodied Visual Navigation
Daniel Gordon, Abhishek Kadian, Devi Parikh, Judy Hoffman, Dhruv Batra; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1022-1031
We propose SplitNet, a method for decoupling visual perception and policy learning. By incorporating auxiliary tasks and selective learning of portions of the model, we explicitly decompose the learning objectives for visual navigation into perceiving the world and acting on that perception. We show improvements over baseline models on transferring between simulators, an encouraging step towards Sim2Real. Additionally, SplitNet generalizes better to unseen environments from t

he same simulator and transfers faster and more effectively to novel embodied na vigation tasks. Further, given only a small sample from a target domain, SplitNe t can match the performance of traditional end-to-end pipelines which receive th e entire dataset

****

Cascaded Parallel Filtering for Memory-Efficient Image-Based Localization
Wentao Cheng, Weisi Lin, Kan Chen, Xinfeng Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1032-1041
Image-based localization (IBL) aims to estimate the 6DOF camera pose for a given query image. The camera pose can be computed from 2D-3D matches between a query image and Structure-from-Motion (SfM) models. Despite recent advances in IBL, i t remains difficult to simultaneously resolve the memory consumption and match a mbiguity problems of large SfM models. In this work, we propose a cascaded paral lel filtering method that leverages the feature, visibility and geometry informa tion to filter wrong matches under binary feature representation. The core idea is that we divide the challenging filtering task into two parallel tasks before deriving an auxiliary camera pose for final filtering. One task focuses on prese rving potentially correct matches, while another focuses on obtaining high quali ty matches to facilitate subsequent more powerful filtering. Moreover, our propo sed method improves the localization accuracy by introducing a quality-aware spa tial reconfiguration method and a principal focal length enhanced pose estimatio n method. Experimental results on real-world datasets demonstrate that our metho d achieves very competitive localization performances in a memory-efficient mann er.

****

Pixel2Mesh++: Multi-View 3D Mesh Generation via Deformation
Chao Wen, Yinda Zhang, Zhuwen Li, Yanwei Fu; Proceedings of the IEEE/CVF Inte rnational Conference on Computer Vision (ICCV), 2019, pp. 1042-1051
We study the problem of shape generation in 3D mesh representation from a few co lor images with known camera poses. While many previous works learn to hallucina te the shape directly from priors, we resort to further improving the shape qual ity by leveraging cross-view information with a graph convolutional network. Ins tead of building a direct mapping function from images to 3D shape, our model le arns to predict series of deformations to improve a coarse shape iteratively. In spired by traditional multiple view geometry methods, our network samples nearby area around the initial mesh's vertex locations and reasons an optimal deformat ion using perceptual feature statistics built from multiple input images. Extens ive experiments show that our model produces accurate 3D shape that are not only visually plausible from the input perspectives, but also well aligned to arbitr ary viewpoints. With the help of physically driven architecture, our model also exhibits generalization capability across different semantic categories, number of input images, and quality of mesh initialization.

****

A Differential Volumetric Approach to Multi-View Photometric Stereo
Fotios Logothetis, Roberto Mecca, Roberto Cipolla; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1052-1061
Highly accurate 3D volumetric reconstruction is still an open research topic whe re the main difficulty is usually related to merging some rough estimations with high frequency details. One of the most promising methods is the fusion between multi-view stereo and photometric stereo images. Beside the intrinsic difficult ies that multi-view stereo and photometric stereo in order to work reliably, sup plementary problems arise when considered together. In this work, we present a v olumetric approach to the multi-view photometric stereo problem. The key point o f our method is the signed distance field parameterisation and its relation to t he surface normal. This is exploited in order to obtain a linear partial differe ntial equation which is solved in a variational framework, that combines multipl e images from multiple points of view in a single system. In addition, the volum etric approach is naturally implemented on an octree, which allows for fast ray-tracing that reliably alleviates occlusions and cast shadows. Our approach is ev aluated on synthetic and real data-sets and achieves state-of-the-art results.

****
Revisiting Radial Distortion Absolute Pose
Viktor Larsson, Torsten Sattler, Zuzana Kukelova, Marc Pollefeys; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1062-1071
To model radial distortion there are two main approaches; either the image points are undistorted such that they correspond to pinhole projections, or the pinhole projections are distorted such that they align with the image measurements. Depending on the application, either of the two approaches can be more suitable. For example, distortion models are commonly used in Structure-from-Motion since they simplify measuring the reprojection error in images. Surprisingly, all previous minimal solvers for pose estimation with radial distortion use undistortion models. In this paper we aim to fill this gap in the literature by proposing the first minimal solvers which can jointly estimate distortion models together with camera pose. We present a general approach which can handle rational models of arbitrary degree for both distortion and undistortion.
****
Estimating the Fundamental Matrix Without Point Correspondences With Application to Transmission Imaging
Tobias Wurfl, Andre Aichert, Nicole Maass, Frank Dennerlein, Andreas Maier; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1072-1081
We present a general method to estimate the fundamental matrix from a pair of images under perspective projection without the need for image point correspondences. Our method is particularly well-suited for transmission imaging, where state-of-the-art feature detection and matching approaches generally do not perform well. Estimation of the fundamental matrix plays a central role in auto-calibration methods for reflection imaging. Such methods are currently not applicable to transmission imaging. Furthermore, our method extends an existing technique proposed for reflection imaging which potentially avoids the outlier-prone feature matching step from an orthographic projection model to a perspective model. Our method exploits the idea that under a linear attenuation model line integrals along corresponding epipolar lines are equal if we compute their derivatives in orthogonal direction to their common epipolar plane. We use the fundamental matrix to parametrize this equality. Our method estimates the matrix by formulating a non-convex optimization problem, minimizing an error in our measurement of this equality. We believe this technique will enable the application of the large body of work on image-based camera pose estimation to transmission imaging leading to more accurate and more general motion compensation and auto-calibration algorithms, particularly in medical X-ray and Computed Tomography imaging.
****
QUARCH: A New Quasi-Affine Reconstruction Stratum From Vague Relative Camera Orientation Knowledge
Devesh Adlakha, Adlane Habed, Fabio Morbidi, Cedric Demonceaux, Michel de Mathelin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1082-1090
We present a new quasi-affine reconstruction of a scene and its application to camera self-calibration. We refer to this reconstruction as QUARCH (QUasi-Affine Reconstruction with respect to Camera centers and the Hodographs of horopters). A QUARCH can be obtained by solving a semidefinite programming problem when, (i) the images have been captured by a moving camera with constant intrinsic parameters, and (ii) a vague knowledge of the relative orientation (under or over 120 degrees) between camera pairs is available. The resulting reconstruction comes close enough to an affine one allowing thus an easy upgrade of the QUARCH to its affine and metric counterparts. We also present a constrained Levenberg-Marquardt method for nonlinear optimization subject to Linear Matrix Inequality (LMI) constraints so as to ensure that the QUARCH LMIs are satisfied during optimization. Experiments with synthetic and real data show the benefits of QUARCH in reliably obtaining a metric reconstruction.
****

Homography From Two Orientation- and Scale-Covariant Features
Daniel Barath, Zuzana Kukelova; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1091-1099
This paper proposes a geometric interpretation of the angles and scales which the orientation- and scale-covariant feature detectors, e.g. SIFT, provide. Two new general constraints are derived on the scales and rotations which can be used in any geometric model estimation tasks. Using these formulas, two new constraints on homography estimation are introduced. Exploiting the derived equations, a solver for estimating the homography from the minimal number of two correspondences is proposed. Also, it is shown how the normalization of the point correspondences affects the rotation and scale parameters, thus achieving numerically stable results. Due to requiring merely two feature pairs, robust estimators, e.g. RANSAC, do significantly fewer iterations than by using the four-point algorithm. When using covariant features, e.g. SIFT, this additional information is given at no cost. The method is tested in a synthetic environment and on publicly available real-world datasets.
****

Hiding Video in Audio via Reversible Generative Models
Hyukryul Yang, Hao Ouyang, Vladlen Koltun, Qifeng Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1100-1109
We present a method for hiding video content inside audio files while preserving the perceptual fidelity of the cover audio. This is a form of cross-modal steganography and is particularly challenging due to the high bitrate of video. Our scheme uses recent advances in flow-based generative models, which enable mapping audio to latent codes such that nearby codes correspond to perceptually similar signals. We show that compressed video data can be concealed in the latent codes of audio sequences while preserving the fidelity of both the hidden video and the cover audio. We can embed 128x128 video inside same-duration audio, or higher-resolution video inside longer audio sequences. Quantitative experiments show that our approach outperforms relevant baselines in steganographic capacity and fidelity.
****

GSLAM: A General SLAM Framework and Benchmark
Yong Zhao, Shibiao Xu, Shuhui Bu, Hongkai Jiang, Pengcheng Han; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1110-1120
SLAM technology has recently seen many successes and attracted the attention of high-technological companies. However, how to unify the interface of existing or emerging algorithms, and effectively perform benchmark about the speed, robustness and portability are still problems. In this paper, we propose a novel SLAM platform named GSLAM, which not only provides evaluation functionality, but also supplies useful toolkit for researchers to quickly develop their SLAM systems. Our core contribution is an universal, cross-platform and full open-source SLAM interface for both research and commercial usage, which is aimed to handle interactions with input dataset, SLAM implementation, visualization and applications in an unified framework. Through this platform, users can implement their own functions for better performance with plugin form and further boost the application to practical usage of the SLAM.
****

Elaborate Monocular Point and Line SLAM With Robust Initialization
Sang Jun Lee, Sung Soo Hwang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1121-1129
This paper presents a monocular indirect SLAM system which performs robust initialization and accurate localization. For initialization, we utilize a matrix factorization-based method. Matrix factorization-based methods require that extracted feature points must be tracked in all used frames. Since consistent tracking is difficult in challenging environments, a geometric interpolation that utilizes epipolar geometry is proposed. For localization, 3D lines are utilized. We propose the use of Plu cker line coordinates to represent geometric information of lines. We also propose orthonormal representation of Plu cker line coordinates a

nd Jacobians of lines for better optimization. Experimental results show that th e proposed initialization generates consistent and robust map in linear time wit h fast convergence even in challenging scenes. And localization using proposed l ine representations is faster, more accurate and memory efficient than other sta te-of-the-art methods.
****

Adaptive Density Map Generation for Crowd Counting
Jia Wan, Antoni Chan; Proceedings of the IEEE/CVF International Conference on C omputer Vision (ICCV), 2019, pp. 1130-1139
Crowd counting is an important topic in computer vision due to its practical usa ge in surveillance systems. The typical design of crowd counting algorithms is d ivided into two steps. First, the ground-truth density maps of crowd images are generated from the ground-truth dot maps (density map generation), e.g., by conv olving with a Gaussian kernel. Second, deep learning models are designed to pred ict a density map from an input image (density map estimation). Most research ef forts have concentrated on the density map estimation problem, while the problem of density map generation has not been adequately explored. In particular, the density map could be considered as an intermediate representation used to train a crowd counting network. In the sense of end-to-end training, the hand-crafted methods used for generating the density maps may not be optimal for the particul ar network or dataset used. To address this issue, we first show the impact of d ifferent density maps and that better ground-truth density maps can be obtained by refining the existing ones using a learned refinement network, which is joint ly trained with the counter. Then, we propose an adaptive density map generator, which takes the annotation dot map as input, and learns a density map represent ation for a counter. The counter and generator are trained jointly within an end -to-end framework. The experiment results on popular counting datasets confirm t he effectiveness of the proposed learnable density map representations.
****

Attention-Aware Polarity Sensitive Embedding for Affective Image Retrieval
Xingxu Yao, Dongyu She, Sicheng Zhao, Jie Liang, Yu-Kun Lai, Jufeng Yang; P roceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2 019, pp. 1140-1150
Images play a crucial role for people to express their opinions online due to th e increasing popularity of social networks. While an affective image retrieval s ystem is useful for obtaining visual contents with desired emotions from a massi ve repository, the abstract and subjective characteristics make the task challen ging. To address the problem, this paper introduces an Attention-aware Polarity Sensitive Embedding (APSE) network to learn affective representations in an end- to-end manner. First, to automatically discover and model the informative region s of interest, we develop a hierarchical attention mechanism, in which both pola rity- and emotion-specific attended representations are aggregated for discrimin ative feature embedding. Second, we present a weighted emotion-pair loss to take the inter- and intra-polarity relationships of the emotional labels into consid eration. Guided by attention module, we weight the sample pairs adaptively which further improves the performance of feature embedding. Extensive experiments on four popular benchmark datasets show that the proposed method performs favorabl y against the state-of-the-art approaches.
****

Zero-Shot Emotion Recognition via Affective Structural Embedding
Chi Zhan, Dongyu She, Sicheng Zhao, Ming-Ming Cheng, Jufeng Yang; Proceeding s of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1151-1160
Image emotion recognition attracts much attention in recent years due to its wid e applications. It aims to classify the emotional response of humans, where cand idate emotion categories are generally defined by specific psychological theorie s, such as Ekman's six basic emotions. However, with the development of psycholo gical theories, emotion categories become increasingly diverse, fine-grained, an d difficult to collect samples. In this paper, we investigate zero-shot learning (ZSL) problem in the emotion recognition task, which tries to recognize the new

unseen emotions. Specifically, we propose a novel affective-structural embeddin g framework, utilizing mid-level semantic representation, i.e., adjective-noun p airs (ANP) features, to construct an affective embedding space. By doing this, t he learned intermediate space can narrow the semantic gap between low-level visu al and high-level semantic features. In addition, we introduce an affective adve rsarial constraint to retain the discriminative capacity of visual features and the affective structural information of semantic features during training proces s. Our method is evaluated on five widely used affective datasets and the perime ntal results show the proposed algorithm outperforms the state-of-the-art approa ches.

****

FW-GAN: Flow-Navigated Warping GAN for Video Virtual Try-On

Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bowen Wu, Bing-Cheng Chen, Jian Yi n; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1161-1170

Beyond current image-based virtual try-on systems that have attracted increasing attention, we move a step forward to developing a video virtual try-on system t hat precisely transfers clothes onto the person and generates visually realistic videos conditioned on arbitrary poses. Besides the challenges in image-based vi rtual try-on (e.g., clothes fidelity, image synthesis), video virtual try-on fur ther requires spatiotemporal consistency. Directly adopting existing image-based approaches often fails to generate coherent video with natural and realistic te xtures. In this work, we propose Flow-navigated Warping Generative Adversarial N etwork (FW-GAN), a novel framework that learns to synthesize the video of virtua l try-on based on a person image, the desired clothes image, and a series of tar get poses. FW-GAN aims to synthesize the coherent and natural video while manipu lating the pose and clothes. It consists of: (i) a flow-guided fusion module tha t warps the past frames to assist synthesis, which is also adopted in the discri minator to help enhance the coherence and quality of the synthesized video; (ii) a warping net that is designed to warp clothes image for the refinement of clot hes textures; (iii) a parsing constraint loss that alleviates the problem caused by the misalignment of segmentation maps from images with different poses and v arious clothes. Experiments on our newly collected dataset show that FW-GAN can synthesize high-quality video of virtual try-on and significantly outperforms ot her methods both qualitatively and quantitatively.

****

Interactive Sketch & Fill: Multiclass Sketch-to-Image Translation

Arnab Ghosh, Richard Zhang, Puneet K. Dokania, Oliver Wang, Alexei A. Efros, Philip H. S. Torr, Eli Shechtman; Proceedings of the IEEE/CVF International C onference on Computer Vision (ICCV), 2019, pp. 1171-1180

We propose an interactive GAN-based sketch-to-image translation method that help s novice users easily create images of simple objects. The user starts with a sp arse sketch and a desired object category, and the network then recommends its p lausible completion(s) and shows a corresponding synthesized image. This enables a feedback loop, where the user can edit the sketch based on the network's reco mmendations, while the network is able to better synthesize the image that the u ser might have in mind. In order to use a single model for a wide array of objec t classes, we introduce a gating-based approach for class conditioning, which al lows us to generate distinct classes without feature mixing, from a single gener ator network.

****

Attention-Based Autism Spectrum Disorder Screening With Privileged Modality

Shi Chen, Qi Zhao; Proceedings of the IEEE/CVF International Conference on Comp uter Vision (ICCV), 2019, pp. 1181-1190

This paper presents a novel framework for automatic and quantitative screening o f autism spectrum disorder (ASD). It is motivated to address two issues in the c urrent clinical settings: 1) short of clinical resources with the prevalence of ASD (1.7% in the United States), and 2) subjectivity of ASD screening. This work differentiates itself with three unique features: first, it proposes an ASD scr eening with privileged modality framework that integrates information from two b

ehavioral modalities during training and improves the performance on each single modality at testing. The proposed framework does not require overlap in subjects between the modalities. Second, it develops the first computational model to classify people with ASD using a photo-taking task where subjects freely explore their environment in a more ecological setting. Photo-taking reveals attentional preference of subjects, differentiating people with ASD from healthy people, and is also easy to implement in real-world clinical settings without requiring advanced diagnostic instruments. Third, this study for the first time takes advantage of the temporal information in eye movements while viewing images, encoding more detailed behavioral differences between ASD people and healthy controls. Experiments show that our ASD screening models can achieve superior performance, outperforming the previous state-of-the-art methods by a considerable margin. Moreover, our framework using diverse modalities demonstrates performance improvement on both the photo-taking and image-viewing tasks, providing a general paradigm that takes in multiple sources of behavioral data for a more accurate ASD screening. The framework is also applicable to various scenarios where one-to-one pairwise relationship is difficult to obtain across different modalities.

****

Image Aesthetic Assessment Based on Pairwise Comparison A Unified Approach to Score Regression, Binary Classification, and Personalization

Jun-Tae Lee, Chang-Su Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1191-1200

We propose a unified approach to three tasks of aesthetic score regression, binary aesthetic classification, and personalized aesthetics. First, we develop a comparator to estimate the ratio of aesthetic scores for two images. Then, we construct a pairwise comparison matrix for multiple reference images and an input image, and predict the aesthetic score of the input via the eigenvalue decomposition of the matrix. By varying the reference images, the proposed algorithm can be used for binary aesthetic classification and personalized aesthetics, as well as generic score regression. Experimental results demonstrate that the proposed unified algorithm provides the state-of-the-art performances in all three tasks of image aesthetics.

****

Delving Into Robust Object Detection From Unmanned Aerial Vehicles: A Deep Nuisance Disentanglement Approach

Zhenyu Wu, Karthik Suresh, Priya Narayanan, Hongyu Xu, Heesung Kwon, Zhangyang Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1201-1210

Object detection from images captured by Unmanned Aerial Vehicles (UAVs) is becoming increasingly useful. Despite the great success of the generic object detection methods trained on ground-to-ground images, a huge performance drop is observed when they are directly applied to images captured by UAVs. The unsatisfactory performance is owing to many UAV-specific nuisances, such as varying flying altitudes, adverse weather conditions, dynamically changing viewing angles, etc. Those nuisances constitute a large number of fine-grained domains, across which the detection model has to stay robust. Fortunately, UAVs will record meta-data that depict those varying attributes, which are either freely available along with the UAV images, or can be easily obtained. We propose to utilize those free meta-data in conjunction with associated UAV images to learn domain-robust features via an adversarial training framework dubbed Nuisance Disentangled Feature Transform (NDFT), for the specific challenging problem of object detection in UAV images, achieving a substantial gain in robustness to those nuisances. We demonstrate the effectiveness of our proposed algorithm, by showing state-of-the- art performance (single model) on two existing UAV-based object detection benchmarks. The code is available at https://github.com/TAMU-VITA/UAV-NDFT.

****

Bit-Flip Attack: Crushing Neural Network With Progressive Bit Search

Adnan Siraj Rakin, Zhezhi He, Deliang Fan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1211-1220

Several important security issues of Deep Neural Network (DNN) have been raised

recently associated with different applications and components. The most widely investigated security concern of DNN is from its malicious input, a.k.a adversarial example. Nevertheless, the security challenge of DNN's parameters is not well explored yet. In this work, we are the first to propose a novel DNN weight attack methodology called Bit-Flip Attack (BFA) which can crush a neural network through maliciously flipping extremely small amount of bits within its weight storage memory system (i.e., DRAM). The bit-flip operations could be conducted through well-known Row-Hammer attack, while our main contribution is to develop an algorithm to identify the most vulnerable bits of DNN weight parameters (stored in memory as binary bits), that could maximize the accuracy degradation with a minimum number of bit-flips. Our proposed BFA utilizes a Progressive Bit Search (PBS) method which combines gradient ranking and progressive search to identify the most vulnerable bit to be flipped. With the aid of PBS, we can successfully attack a ResNet-18 fully malfunction (i.e., top-1 accuracy degrade from 69.8% to 0.1%) only through 13 bit-flips out of 93 million bits, while randomly flipping 100 bits merely degrades the accuracy by less than 1%. Code is released at: https://github.com/elliothe/Neural_Network_Weight_Attack

****

Employing Deep Part-Object Relationships for Salient Object Detection

Yi Liu, Qiang Zhang, Dingwen Zhang, Jungong Han; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1232-1241

Despite Convolutional Neural Networks (CNNs) based methods have been successful in detecting salient objects, their underlying mechanism that decides the salient intensity of each image part separately cannot avoid inconsistency of parts within the same salient object. This would ultimately result in an incomplete shape of the detected salient object. To solve this problem, we dig into part-object relationships and take the unprecedented attempt to employ these relationships endowed by the Capsule Network (CapsNet) for salient object detection. The entire salient object detection system is built directly on a Two-Stream Part-Object Assignment Network (TSPOANet) consisting of three algorithmic steps. In the first step, the learned deep feature maps of the input image are transformed to a group of primary capsules. In the second step, we feed the primary capsules into two identical streams, within each of which low-level capsules (parts) will be assigned to their familiar high-level capsules (object) via a locally connected routing. In the final step, the two streams are integrated in the form of a fully connected layer, where the relevant parts can be clustered together to form a complete salient object. Experimental results demonstrate the superiority of the proposed salient object detection network over the state-of-the-art methods.

****

Self-Supervised Deep Depth Denoising

Vladimiros Sterzentsenko, Leonidas Saroglou, Anargyros Chatzitofis, Spyridon Thermos, Nikolaos Zioulis, Alexandros Doumanoglou, Dimitrios Zarpalas, Petros Daras; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1242-1251

Depth perception is considered an invaluable source of information for various vision tasks. However, depth maps acquired using consumer-level sensors still suffer from non-negligible noise. This fact has recently motivated researchers to exploit traditional filters, as well as the deep learning paradigm, in order to suppress the aforementioned non-uniform noise, while preserving geometric details. Despite the effort, deep depth denoising is still an open challenge mainly due to the lack of clean data that could be used as ground truth. In this paper, we propose a fully convolutional deep autoencoder that learns to denoise depth maps, surpassing the lack of ground truth data. Specifically, the proposed autoencoder exploits multiple views of the same scene from different points of view in order to learn to suppress noise in a self-supervised end-to-end manner using depth and color information during training, yet only depth during inference. To enforce self-supervision, we leverage a differentiable rendering technique to exploit photometric supervision, which is further regularized using geometric and surface priors. As the proposed approach relies on raw data acquisition, a large RGB-D corpus is collected using Intel RealSense sensors. Complementary to a quant

itative evaluation, we demonstrate the effectiveness of the proposed self-supervised denoising approach on established 3D reconstruction applications. Code is available at https://github.com/VCL3D/DeepDepthDenoising

****

Cost-Aware Fine-Grained Recognition for IoTs Based on Sequential Fixations
Hanxiao Wang, Venkatesh Saligrama, Stan Sclaroff, Vitaly Ablavsky; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1252-1261

We consider the problem of fine-grained classification on an edge camera device that has limited power. The edge device must sparingly interact with the cloud to minimize communication bits to conserve power, and the cloud upon receiving the edge inputs returns a classification label. To deal with fine-grained classification, we adopt the perspective of sequential fixation with a foveated field-of-view to model cloud-edge interactions. We propose a novel deep reinforcement learning-based foveation model, DRIFT, that sequentially generates and recognizes mixed-acuity images. Training of DRIFT requires only image-level category labels and encourages fixations to contain task-relevant information, while maintaining data efficiency. Specifically, we train a foveation actor network with a novel Deep Deterministic Policy Gradient by Conditioned Critic and Coaching(DDPGC3) algorithm. In addition, we propose to shape the reward to provide informative feedback after each fixation to better guide RL training. We demonstrate the effectiveness of DRIFT on this task by evaluating on five fine-grained classification benchmark datasets, and show that the proposed approach achieves state-of-the-art performance with over 3X reduction in transmitted pixels.

****

Layout-Induced Video Representation for Recognizing Agent-in-Place Actions
Ruichi Yu, Hongcheng Wang, Ang Li, Jingxiao Zheng, Vlad I. Morariu, Larry S. Davis; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1262-1272

We address scene layout modeling for recognizing agent-in-place actions, which are actions associated with agents who perform them and the places where they occur, in the context of outdoor home surveillance. We introduce a novel representation to model the geometry and topology of scene layouts so that a network can generalize from the layouts observed in the training scenes to unseen scenes in the test set. This Layout-Induced Video Representation (LIVR) abstracts away low-level appearance variance and encodes geometric and topological relationships of places to explicitly model scene layout. LIVR partitions the semantic features of a scene into different places to force the network to learn generic place-based feature descriptions which are independent of specific scene layouts; then, LIVR dynamically aggregates features based on connectivities of places in each specific scene to model its layout. We introduce a new Agent-in-Place Action (APA) dataset(The dataset is pending legal review and will be released upon the acceptance of this paper.) to show that our method allows neural network models to generalize significantly better to unseen scenes.

****

Anomaly Detection in Video Sequence With Appearance-Motion Correspondence
Trong-Nguyen Nguyen, Jean Meunier; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1273-1283

Anomaly detection in surveillance videos is currently a challenge because of the diversity of possible events. We propose a deep convolutional neural network (CNN) that addresses this problem by learning a correspondence between common object appearances (e.g. pedestrian, background, tree, etc.) and their associated motions. Our model is designed as a combination of a reconstruction network and an image translation model that share the same encoder. The former sub-network determines the most significant structures that appear in video frames and the latter one attempts to associate motion templates to such structures. The training stage is performed using only videos of normal events and the model is then capable to estimate frame-level scores for an unknown input. The experiments on 6 benchmark datasets demonstrate the competitive performance of the proposed approach with respect to state-of-the-art methods.

****
Exploring Randomly Wired Neural Networks for Image Recognition
Saining Xie, Alexander Kirillov, Ross Girshick, Kaiming He; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1284-1293
Neural networks for image recognition have evolved through extensive manual design from simple chain-like models to structures with multiple wiring paths. The success of ResNets and DenseNets is due in large part to their innovative wiring plans. Now, neural architecture search (NAS) studies are exploring the joint optimization of wiring and operation types, however, the space of possible wirings is constrained and still driven by manual design despite being searched. In this paper, we explore a more diverse set of connectivity patterns through the lens of randomly wired neural networks. To do this, we first define the concept of a stochastic network generator that encapsulates the entire network generation process. Encapsulation provides a unified view of NAS and randomly wired networks. Then, we use three classical random graph models to generate randomly wired graphs for networks. The results are surprising: several variants of these random generators yield network instances that have competitive accuracy on the ImageNet benchmark. These results suggest that new efforts focusing on designing better network generators may lead to new breakthroughs by exploring less constrained search spaces with more room for novel design.
****
Progressive Differentiable Architecture Search: Bridging the Depth Gap Between Search and Evaluation
Xin Chen, Lingxi Xie, Jun Wu, Qi Tian; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1294-1303
Recently, differentiable search methods have made major progress in reducing the computational costs of neural architecture search. However, these approaches often report lower accuracy in evaluating the searched architecture or transferring it to another dataset. This is arguably due to the large gap between the architecture depths in search and evaluation scenarios. In this paper, we present an efficient algorithm which allows the depth of searched architectures to grow gradually during the training procedure. This brings two issues, namely, heavier computational overheads and weaker search stability, which we solve using search space approximation and regularization, respectively. With a significantly reduced search time (7 hours on a single GPU), our approach achieves state-of-the-art performance on both the proxy dataset (CIFAR10 or CIFAR100) and the target dataset (ImageNet). Code is available at https://github.com/chenxin061/pdarts
****
Multinomial Distribution Learning for Effective Neural Architecture Search
Xiawu Zheng, Rongrong Ji, Lang Tang, Baochang Zhang, Jianzhuang Liu, Qi Tian; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1304-1313
Architectures obtained by Neural Architecture Search (NAS) have achieved highly competitive performance in various computer vision tasks. However, the prohibitive computation demand of forward-backward propagation in deep neural networks and searching algorithms makes it difficult to apply NAS in practice. In this paper, we propose a Multinomial Distribution Learning for extremely effective NAS, which considers the search space as a joint multinomial distribution, i.e., the operation between two nodes is sampled from this distribution, and the optimal network structure is obtained by the operations with the most likely probability in this distribution. Therefore, NAS can be transformed to a multinomial distribution learning problem, i.e., the distribution is optimized to have a high expectation of the performance. Besides, a hypothesis that the performance ranking is consistent in every training epoch is proposed and demonstrated to further accelerate the learning process. Experiments on CIFAR-10 and ImageNet demonstrate the effectiveness of our method. On CIFAR-10, the structure searched by our method achieves 2.55% test error, while being 6.0x (only 4 GPU hours on GTX1080Ti) faster compared with state-of-the-art NAS algorithms. On ImageNet, our model achieves 74% top1 accuracy under MobileNet settings (MobileNet V1/V2), while being 1.2x

faster with measured GPU latency. Test code with pre-trained models are availab
le at https: //github.com/tanglang96/MDENAS
****

Searching for MobileNetV3
Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing
Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, H
artwig Adam; Proceedings of the IEEE/CVF International Conference on Computer Vi
sion (ICCV), 2019, pp. 1314-1324
We present the next generation of MobileNets based on a combination of complemen
tary search techniques as well as a novel architecture design. MobileNetV3 is tu
ned to mobile phone CPUs through a combination of hardware-aware network archite
cture search (NAS) complemented by the NetAdapt algorithm and then subsequently
improved through novel architecture advances. This paper starts the exploration
of how automated search algorithms and network design can work together to harne
ss complementary approaches improving the overall state of the art. Through this
process we create two new MobileNet models for release: MobileNetV3-Large and M
obileNetV3-Small which are targeted for high and low resource use cases. These m
odels are then adapted and applied to the tasks of object detection and semantic
segmentation. For the task of semantic segmentation (or any dense pixel predict
ion), we propose a new efficient segmentation decoder Lite Reduced Atrous Spatia
l Pyramid Pooling (LR-ASPP). We achieve new state of the art results for mobile
classification, detection and segmentation. MobileNetV3-Large is 3.2% more accur
ate on ImageNet classification while reducing latency by 20% compared to MobileN
etV2. MobileNetV3-Small is 6.6% more accurate compared to a MobileNetV2 model wi
th comparable latency. MobileNetV3-Large detection is over 25% faster at roughly
the same accuracy as MobileNetV2 on COCO detection. MobileNetV3-Large LR-ASPP i
s 34% faster than MobileNetV2 R-ASPP at similar accuracy for Cityscapes segmenta
tion.
****

Data-Free Quantization Through Weight Equalization and Bias Correction
Markus Nagel, Mart van Baalen, Tijmen Blankevoort, Max Welling; Proceedings o
f the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 132
5-1334
We introduce a data-free quantization method for deep neural networks that does
not require fine-tuning or hyperparameter selection. It achieves near-original m
odel performance on common computer vision architectures and tasks. 8-bit fixed-
point quantization is essential for efficient inference on modern deep learning
hardware. However, quantizing models to run in 8-bit is a non-trivial task, freq
uently leading to either significant performance reduction or engineering time s
pent on training a network to be amenable to quantization. Our approach relies o
n equalizing the weight ranges in the network by making use of a scale-equivaria
nce property of activation functions. In addition the method corrects biases in
the error that are introduced during quantization. This improves quantization ac
curacy performance, and can be applied to many common computer vision architectu
res with a straight forward API call. For common architectures, such as the Mobi
leNet family, we achieve state-of-the-art quantized model performance. We furthe
r show that the method also extends to other computer vision architectures and t
asks such as semantic segmentation and object detection.
****

A Camera That CNNs: Towards Embedded Neural Networks on Pixel Processor Arrays
Laurie Bose, Jianing Chen, Stephen J. Carey, Piotr Dudek, Walterio Mayol-Cue
vas; Proceedings of the IEEE/CVF International Conference on Computer Vision (IC
CV), 2019, pp. 1335-1344
We present a convolutional neural network implementation for pixel processor arr
ay (PPA) sensors. PPA hardware consists of a fine-grained array of general-purpo
se processing elements, each capable of light capture, data storage, program exe
cution, and communication with neighboring elements. This allows images to be st
ored and manipulated directly at the point of light capture, rather than having
to transfer images to external processing hardware. Our CNN approach divides thi
s array up into 4x4 blocks of processing elements, essentially trading-off image

resolution for increased local memory capacity per 4x4 "pixel". We implement pa
rallel operations for image addition, subtraction and bit-shifting images in thi
s 4x4 block format. Using these components we formulate how to perform ternary w
eight convolutions upon these images, compactly store results of such convolutio
ns, perform max-pooling, and transfer the resulting sub-sampled data to an attac
hed micro-controller. We train ternary weight filter CNNs for digit recognition
and a simple tracking task, and demonstrate inference of these networks upon the
SCAMP5 PPA system. This work represents a first step towards embedding neural n
etwork processing capability directly onto the focal plane of a sensor.
****

Knowledge Distillation via Route Constrained Optimization
Xiao Jin, Baoyun Peng, Yichao Wu, Yu Liu, Jiaheng Liu, Ding Liang, Junjie
Yan, Xiaolin Hu; Proceedings of the IEEE/CVF International Conference on Comput
er Vision (ICCV), 2019, pp. 1345-1354
Distillation-based learning boosts the performance of the miniaturized neural ne
twork based on the hypothesis that the representation of a teacher model can be
used as structured and relatively weak supervision, and thus would be easily lea
rned by a miniaturized model. However, we find that the representation of a conv
erged heavy model is still a strong constraint for training a small student mode
l, which leads to a higher lower bound of congruence loss. In this work, we cons
ider the knowledge distillation from the perspective of curriculum learning by t
eacher's routing. Instead of supervising the student model with a converged teac
her model, we supervised it with some anchor points selected from the route in p
arameter space that the teacher model passed by, as we called route constrained
optimization (RCO). We experimentally demonstrate this simple operation greatly
reduces the lower bound of congruence loss for knowledge distillation, hint and
mimicking learning. On close-set classification tasks like CIFAR and ImageNet, R
CO improves knowledge distillation by 2.14% and 1.5% respectively. For the sake
of evaluating the generalization, we also test RCO on the open-set face recognit
ion task MegaFace. RCO achieves 84.3% accuracy on one-to-million task with only
0.8 M parameters, which push the SOTA by a large margin.
****

Distillation-Based Training for Multi-Exit Architectures
Mary Phuong, Christoph H. Lampert; Proceedings of the IEEE/CVF International Co
nference on Computer Vision (ICCV), 2019, pp. 1355-1364
Multi-exit architectures, in which a stack of processing layers is interleaved w
ith early output layers, allow the processing of a test example to stop early an
d thus save computation time and/or energy. In this work, we propose a new train
ing procedure for multi-exit architectures based on the principle of knowledge d
istillation. The method encourages early exits to mimic later, more accurate exi
ts, by matching their probability outputs. Experiments on CIFAR100 and ImageNet
show that distillation-based training significantly improves the accuracy of ear
ly exits while maintaining state-of-the-art accuracy for late ones. The method i
s particularly beneficial when training data is limited and also allows a straig
ht-forward extension to semi-supervised learning, i.e. make use also of unlabele
d data at training time. Moreover, it takes only a few lines to implement and im
poses almost no computational overhead at training time, and none at all at test
time.
****

Similarity-Preserving Knowledge Distillation
Frederick Tung, Greg Mori; Proceedings of the IEEE/CVF International Conference
on Computer Vision (ICCV), 2019, pp. 1365-1374
Knowledge distillation is a widely applicable technique for training a student n
eural network under the guidance of a trained teacher network. For example, in n
eural network compression, a high-capacity teacher is distilled to train a compa
ct student; in privileged learning, a teacher trained with privileged data is di
stilled to train a student without access to that data. The distillation loss de
termines how a teacher's knowledge is captured and transferred to the student. I
n this paper, we propose a new form of knowledge distillation loss that is inspi
red by the observation that semantically similar inputs tend to elicit similar a

ctivation patterns in a trained network. Similarity-preserving knowledge distill
ation guides the training of a student network such that input pairs that produc
e similar (dissimilar) activations in the teacher network produce similar (dissi
milar) activations in the student network. In contrast to previous distillation
methods, the student is not required to mimic the representation space of the te
acher, but rather to preserve the pairwise similarities in its own representatio
n space. Experiments on three public datasets demonstrate the potential of our a
pproach.
****

Many Task Learning With Task Routing
Gjorgji Strezoski, Nanne van Noord, Marcel Worring; Proceedings of the IEEE/CV
F International Conference on Computer Vision (ICCV), 2019, pp. 1375-1384
Typical multi-task learning (MTL) methods rely on architectural adjustments and
a large trainable parameter set to jointly optimize over several tasks. However,
when the number of tasks increases so do the complexity of the architectural ad
justments and resource requirements. In this paper, we introduce a method which
applies a conditional feature-wise transformation over the convolutional activat
ions that enables a model to successfully perform a large number of tasks. To di
stinguish from regular MTL, we introduce Many Task Learning (MaTL) as a special
case of MTL where more than 20 tasks are performed by a single model. Our method
dubbed Task Routing (TR) is encapsulated in a layer we call the Task Routing La
yer (TRL), which applied in an MaTL scenario successfully fits hundreds of class
ification tasks in one model. We evaluate on 5 datasets and the Visual Decathlon
(VD) challenge against strong baselines and state-of-the-art approaches.
****

Stochastic Filter Groups for Multi-Task CNNs: Learning Specialist and Generalist
Convolution Kernels
Felix J.S. Bragman, Ryutaro Tanno, Sebastien Ourselin, Daniel C. Alexander,
Jorge Cardoso; Proceedings of the IEEE/CVF International Conference on Computer
Vision (ICCV), 2019, pp. 1385-1394
The performance of multi-task learning in Convolutional Neural Networks (CNNs) h
inges on the design of feature sharing between tasks within the architecture. Th
e number of possible sharing patterns are combinatorial in the depth of the netw
ork and the number of tasks, and thus hand-crafting an architecture, purely base
d on the human intuitions of task relationships can be time-consuming and subopt
imal. In this paper, we present a probabilistic approach to learning task-specif
ic and shared representations in CNNs for multi-task learning. Specifically, we
propose "stochastic filter groups" (SFG), a mechanism to assign convolution kern
els in each layer to "specialist" and "generalist" groups, which are specific to
and shared across different tasks, respectively. The SFG modules determine the
connectivity between layers and the structures of task-specific and shared repre
sentations in the network. We employ variational inference to learn the posterio
r distribution over the possible grouping of kernels and network parameters. Exp
eriments demonstrate the proposed method generalises across multiple tasks and s
hows improved performance over baseline methods.
****

Transferability and Hardness of Supervised Classification Tasks
Anh T. Tran, Cuong V. Nguyen, Tal Hassner; Proceedings of the IEEE/CVF Interna
tional Conference on Computer Vision (ICCV), 2019, pp. 1395-1405
We propose a novel approach for estimating the difficulty and transferability of
supervised classification tasks. Unlike previous work, our approach is solution
agnostic and does not require or assume trained models. Instead, we estimate th
ese values using an information theoretic approach: treating training labels as
random variables and exploring their statistics. When transferring from a source
to a target task, we consider the conditional entropy between two such variable
s (i.e., label assignments of the two tasks). We show analytically and empirical
ly that this value is related to the loss of the transferred model. We further s
how how to use this value to estimate task hardness. We test our claims extensiv
ely on three large scale data sets---CelebA (40 tasks), Animals with Attributes
2 (85 tasks), and Caltech-UCSD Birds 200 (312 tasks)---together representing 437

classification tasks. We provide results showing that our hardness and transfer ability estimates are strongly correlated with empirical hardness and transferab ility. As a case study, we transfer a learned face recognition model to CelebA a ttribute classification tasks, showing state of the art accuracy for highly tran sferable attributes.

****

Moment Matching for Multi-Source Domain Adaptation

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, Bo Wang; Pro ceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 201 9, pp. 1406-1415

Conventional unsupervised domain adaptation (UDA) assumes that training data are sampled from a single domain. This neglects the more practical scenario where t raining data are collected from multiple sources, requiring multi-source domain adaptation. We make three major contributions towards addressing this problem. F irst, we collect and annotate by far the largest UDA dataset, called DomainNet, which contains six domains and about 0.6 million images distributed among 345 ca tegories, addressing the gap in data availability for multi-source UDA research. Second, we propose a new deep learning approach, Moment Matching for Multi-Sour ce Domain Adaptation (M3SDA), which aims to transfer knowledge learned from mult iple labeled source domains to an unlabeled target domain by dynamically alignin g moments of their feature distributions. Third, we provide new theoretical insi ghts specifically for moment matching approaches in both single and multiple sou rce domain adaptation. Extensive experiments are conducted to demonstrate the po wer of our new dataset in benchmarking state-of-the-art multi-source domain adap tation methods, as well as the advantage of our proposed model. Dataset and Code are available at http://ai.bu.edu/M3SDA/

****

Unsupervised Domain Adaptation via Regularized Conditional Alignment

Safa Cicek, Stefano Soatto; Proceedings of the IEEE/CVF International Conferenc e on Computer Vision (ICCV), 2019, pp. 1416-1425

We propose a method for unsupervised domain adaptation that trains a shared embe dding to align the joint distributions of inputs (domain) and outputs (classes), making any classifier agnostic to the domain. Joint alignment ensures that not only the marginal distributions of the domains are aligned, but the labels as we ll. We propose a novel objective function that encourages the class-conditional distributions to have disjoint support in feature space. We further exploit adve rsarial regularization to improve the performance of the classifier on the domai n for which no annotated data is available.

****

Larger Norm More Transferable: An Adaptive Feature Norm Approach for Unsupervise d Domain Adaptation

Ruijia Xu, Guanbin Li, Jihan Yang, Liang Lin; Proceedings of the IEEE/CVF Int ernational Conference on Computer Vision (ICCV), 2019, pp. 1426-1435

Domain adaptation enables the learner to safely generalize into novel environmen ts by mitigating domain shifts across distributions. Previous works may not effe ctively uncover the underlying reasons that would lead to the drastic model degr adation on the target task. In this paper, we empirically reveal that the errati c discrimination of the target domain mainly stems from its much smaller feature norms with respect to that of the source domain. To this end, we propose a nove l parameter-free Adaptive Feature Norm approach. We demonstrate that progressive ly adapting the feature norms of the two domains to a large range of values can result in significant transfer gains, implying that those task-specific features with larger norms are more transferable. Our method successfully unifies the co mputation of both standard and partial domain adaptation with more robustness ag ainst the negative transfer issue. Without bells and whistles but a few lines of code, our method substantially lifts the performance on the target task and exc eeds state-of-the-arts by a large margin (11.5% on Office-Home and 17.1% on VisD A2017). We hope our simple yet effective approach will shed some light on the fu ture research of transfer learning. Code is available at https://github.com/jiha nyang/AFN.

****
UM-Adapt: Unsupervised Multi-Task Adaptation Using Adversarial Cross-Task Distil lation
Jogendra Nath Kundu, Nishank Lakkakula, R. Venkatesh Babu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1436-1445
Aiming towards human-level generalization, there is a need to explore adaptable representation learning methods with greater transferability. Most existing appr oaches independently address task-transferability and cross-domain adaptation, r esulting in limited generalization. In this paper, we propose UM-Adapt - a unifi ed framework to effectively perform unsupervised domain adaptation for spatially -structured prediction tasks, simultaneously maintaining a balanced performance across individual tasks in a multi-task setting. To realize this, we propose two novel regularization strategies; a) Contour-based content regularization (CCR) and b) exploitation of inter-task coherency using a cross-task distillation modu le. Furthermore, avoiding a conventional ad-hoc domain discriminator, we re-util ize the cross-task distillation loss as output of an energy function to adversar ially minimize the input domain discrepancy. Through extensive experiments, we d emonstrate superior generalizability of the learned representations simultaneous ly for multiple tasks under domain-shifts from synthetic to natural environments . UM-Adapt yields state-of-the-art transfer learning results on ImageNet classif ication and comparable performance on PASCAL VOC 2007 detection task, even with a smaller backbone-net. Moreover, the resulting semi-supervised framework outper forms the current fully-supervised multi-task learning state-of-the-art on both NYUD and Cityscapes dataset.
****
Episodic Training for Domain Generalization
Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, Timothy M. Hospe dales; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1446-1455
Domain generalization (DG) is the challenging and topical problem of learning mo dels that generalize to novel testing domains with different statistics than a s et of known training domains. The simple approach of aggregating data from all s ource domains and training a single deep neural network end-to-end on all the da ta provides a surprisingly strong baseline that surpasses many prior published m ethods. In this paper we build on this strong baseline by designing an episodic training procedure that trains a single deep network in a way that exposes it to the domain shift that characterises a novel domain at runtime. Specifically, we decompose a deep network into feature extractor and classifier components, and then train each component by simulating it interacting with a partner who is bad ly tuned for the current domain. This makes both components more robust, ultimat ely leading to our networks producing state-of-the-art performance on three DG b enchmarks. Furthermore, we consider the pervasive workflow of using an ImageNet trained CNN as a fixed feature extractor for downstream recognition tasks. Using the Visual Decathlon benchmark, we demonstrate that our episodic-DG training im proves the performance of such a general purpose feature extractor by explicitly training a feature for robustness to novel problems. This shows that DG trainin g can benefit standard practice in computer vision.
****
Domain Adaptation for Structured Output via Discriminative Patch Representations
Yi-Hsuan Tsai, Kihyuk Sohn, Samuel Schulter, Manmohan Chandraker; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1 456-1465
Predicting structured outputs such as semantic segmentation relies on expensive per-pixel annotations to learn supervised models like convolutional neural netwo rks. However, models trained on one data domain may not generalize well to other domains without annotations for model finetuning. To avoid the labor-intensive process of annotation, we develop a domain adaptation method to adapt the source data to the unlabeled target domain. We propose to learn discriminative feature representations of patches in the source domain by discovering multiple modes o f patch-wise output distribution through the construction of a clustered space.

With such representations as guidance, we use an adversarial learning scheme to push the feature representations of target patches in the clustered space closer to the distributions of source patches. In addition, we show that our framework is complementary to existing domain adaptation techniques and achieves consistent improvements on semantic segmentation. Extensive ablations and results are demonstrated on numerous benchmark datasets with various settings, such as synthetic-to-real and cross-city scenarios.
****
Semi-Supervised Learning by Augmented Distribution Alignment
Qin Wang, Wen Li, Luc Van Gool; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1466-1475
In this work, we propose a simple yet effective semi-supervised learning approach called Augmented Distribution Alignment. We reveal that an essential sampling bias exists in semi-supervised learning due to the limited number of labeled samples, which often leads to a considerable empirical distribution mismatch between labeled data and unlabeled data. To this end, we propose to align the empirical distributions of labeled and unlabeled data to alleviate the bias. On one hand, we adopt an adversarial training strategy to minimize the distribution distance between labeled and unlabeled data as inspired by domain adaptation works. On the other hand, to deal with the small sample size issue of labeled data, we also propose a simple interpolation strategy to generate pseudo training samples. Those two strategies can be easily implemented into existing deep neural networks. We demonstrate the effectiveness of our proposed approach on the benchmark SVHN and CIFAR10 datasets. Our code is available at https://github.com/qinenergy/adanet .
****
S4L: Self-Supervised Semi-Supervised Learning
Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, Lucas Beyer; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1476-1485
This work tackles the problem of semi-supervised learning of image classifiers. Our main insight is that the field of semi-supervised learning can benefit from the quickly advancing field of self-supervised visual representation learning. Unifying these two approaches, we propose the framework of self-supervised semi-supervised learning (S4L) and use it to derive two novel semi-supervised image classification methods. We demonstrate the effectiveness of these methods in comparison to both carefully tuned baselines, and existing semi-supervised learning methods. We then show that S4L and existing semi-supervised methods can be jointly trained, yielding a new state-of-the-art result on semi-supervised ILSVRC-2012 with 10% of labels.
****
Privacy Preserving Image Queries for Camera Localization
Pablo Speciale, Johannes L. Schonberger, Sudipta N. Sinha, Marc Pollefeys; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1486-1496
Augmented/mixed reality and robotic applications are increasingly relying on cloud-based localization services, which require users to upload query images to perform camera pose estimation on a server. This raises significant privacy concerns when consumers use such services in their homes or in confidential industrial settings. Even if only image features are uploaded, the privacy concerns remain as the images can be reconstructed fairly well from feature locations and descriptors. We propose to conceal the content of the query images from an adversary on the server or a man-in-the-middle intruder. The key insight is to replace the 2D image feature points in the query image with randomly oriented 2D lines passing through their original 2D positions. It will be shown that this feature representation hides the image contents, and thereby protects user privacy, yet still provides sufficient geometric constraints to enable robust and accurate 6-DOF camera pose estimation from feature correspondences. Our proposed method can handle single- and multi-image queries as well as exploit additional information about known structure, gravity, and scale. Numerous experiments demonstrate the hi

gh practical relevance of our approach.
****

Calibration Wizard: A Guidance System for Camera Calibration Based on Modelling Geometric and Corner Uncertainty
Songyou Peng, Peter Sturm; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1497-1505
It is well known that the accuracy of a calibration depends strongly on the choice of camera poses from which images of a calibration object are acquired. We present a system -- Calibration Wizard -- that interactively guides a user towards taking optimal calibration images. For each new image to be taken, the system computes, from all previously acquired images, the pose that leads to the globally maximum reduction of expected uncertainty on intrinsic parameters and then guides the user towards that pose. We also show how to incorporate uncertainty in corner point position in a novel principled manner, for both, calibration and computation of the next best pose. Synthetic and real-world experiments are performed to demonstrate the effectiveness of Calibration Wizard.
****

Gated2Depth: Real-Time Dense Lidar From Gated Images
Tobias Gruber, Frank Julca-Aguilar, Mario Bijelic, Felix Heide; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1506-1516
We present an imaging framework which converts three images from a gated camera into high-resolution depth maps with depth accuracy comparable to pulsed lidar measurements. Existing scanning lidar systems achieve low spatial resolution at large ranges due to mechanically-limited angular sampling rates, restricting scene understanding tasks to close-range clusters with dense sampling. Moreover, today's pulsed lidar scanners suffer from high cost, power consumption, large form-factors, and they fail in the presence of strong backscatter. We depart from point scanning and demonstrate that it is possible to turn a low-cost CMOS gated imager into a dense depth camera with at least 80m range - by learning depth from three gated images. The proposed architecture exploits semantic context across gated slices, and is trained on a synthetic discriminator loss without the need of dense depth labels. The proposed replacement for scanning lidar systems is real-time, handles back-scatter and provides dense depth at long ranges. We validate our approach in simulation and on real-world data acquired over 4,000km driving in northern Europe. Data and code are available at https://github.com/gruberto/Gated2Depth.
****

X-Section: Cross-Section Prediction for Enhanced RGB-D Fusion
Andrea Nicastro, Ronald Clark, Stefan Leutenegger; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1517-1526
Detailed 3D reconstruction is an important challenge with application to robotics, augmented and virtual reality, which has seen impressive progress throughout the past years. Advancements were driven by the availability of depth cameras (RGB-D), as well as increased compute power, e.g. in the form of GPUs -- but also thanks to inclusion of machine learning in the process. Here, we propose X-Section, an RGB-D 3D reconstruction approach that leverages deep learning to make object-level predictions about thicknesses that can be readily integrated into a volumetric multi-view fusion process, where we propose an extension to the popular KinectFusion approach. In essence, our method allows to complete shape in general indoor scenes behind what is sensed by the RGB-D camera, which may be crucial e.g. for robotic manipulation tasks or efficient scene exploration. Predicting object thicknesses rather than volumes allows us to work with comparably high spatial resolution without exploding memory and training data requirements on the employed Convolutional Neural Networks. In a series of qualitative and quantitative evaluations, we demonstrate how we accurately predict object thickness and reconstruct general 3D scenes containing multiple objects.
****

Learning an Event Sequence Embedding for Dense Event-Based Deep Stereo
Stepan Tulyakov, Francois Fleuret, Martin Kiefel, Peter Gehler, Michael Hirs

ch; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1527-1537
Today, a frame-based camera is the sensor of choice for machine vision applications. However, these cameras, originally developed for acquisition of static images rather than for sensing of dynamic uncontrolled visual environments, suffer from high power consumption, data rate, latency and low dynamic range. An event-based image sensor addresses these drawbacks by mimicking a biological retina. Instead of measuring the intensity of every pixel in a fixed time-interval, it reports events of significant pixel intensity changes. Every such event is represented by its position, sign of change, and timestamp, accurate to the microsecond. Asynchronous event sequences require special handling, since traditional algorithms work only with synchronous, spatially gridded data. To address this problem we introduce a new module for event sequence embedding, for use in difference applications. The module builds a representation of an event sequence by firstly aggregating information locally across time, using a novel fully-connected layer for an irregularly sampled continuous domain, and then across discrete spatial domain. Based on this module, we design a deep learning-based stereo method for event-based cameras. The proposed method is the first learning-based stereo method for an event-based camera and the only method that produces dense results. We show that large performance increases on the Multi Vehicle Stereo Event Camera Dataset (MVSEC), which became the standard set for benchmarking of event-based stereo methods.
****

Point-Based Multi-View Stereo Network
Rui Chen, Songfang Han, Jing Xu, Hao Su; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1538-1547
We introduce Point-MVSNet, a novel point-based deep framework for multi-view stereo (MVS). Distinct from existing cost volume approaches, our method directly processes the target scene as point clouds. More specifically, our method predicts the depth in a coarse-to-fine manner. We first generate a coarse depth map, convert it into a point cloud and refine the point cloud iteratively by estimating the residual between the depth of the current iteration and that of the ground truth. Our network leverages 3D geometry priors and 2D texture information jointly and effectively by fusing them into a feature-augmented point cloud, and processes the point cloud to estimate the 3D flow for each point. This point-based architecture allows higher accuracy, more computational efficiency and more flexibility than cost-volume-based counterparts. Experimental results show that our approach achieves a significant improvement in reconstruction quality compared with state-of-the-art methods on the DTU and the Tanks and Temples dataset. Our source code and trained models are available at https://github.com/callmeray/PointMVSNet.
****

Discrete Laplace Operator Estimation for Dynamic 3D Reconstruction
Xiangyu Xu, Enrique Dunn; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1548-1557
We present a general paradigm for dynamic 3D reconstruction from multiple independent and uncontrolled image sources having arbitrary temporal sampling density and distribution. Our graph-theoretic formulation models the spatio-temporal relationships among our observations in terms of the joint estimation of their 3D geometry and its discrete Laplace operator. Towards this end, we define a tri-convex optimization framework that leverages the geometric properties and dependencies found among a Euclidean shape-space and the discrete Laplace operator describing its local and global topology. We present a reconstructability analysis, experiments on motion capture data and multi-view image datasets, as well as explore applications to geometry-based event segmentation and data association.
****

Deep Non-Rigid Structure From Motion
Chen Kong, Simon Lucey; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1558-1567
Current non-rigid structure from motion (NRSfM) algorithms are mainly limited wi

th respect to: (i) the number of images, and (ii) the type of shape variability they can handle. This has hampered the practical utility of NRSfM for many applications within vision. In this paper we propose a novel deep neural network to recover camera poses and 3D points solely from an ensemble of 2D image coordinates. The proposed neural network is mathematically interpretable as a multi-layer block sparse dictionary learning problem, and can handle problems of unprecedented scale and shape complexity. Extensive experiments demonstrate the impressive performance of our approach where we exhibit superior precision and robustness against all available state-of-the-art works in the order of magnitude. We further propose a quality measure (based on the network weights) which circumvents the need for 3D ground-truth to ascertain the confidence we have in the reconstruction.

****

Equivariant Multi-View Networks

Carlos Esteves, Yinshuang Xu, Christine Allen-Blanchette, Kostas Daniilidis; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1568-1577

Several popular approaches to 3D vision tasks process multiple views of the input independently with deep neural networks pre-trained on natural images, where view permutation invariance is achieved through a single round of pooling over all views. We argue that this operation discards important information and leads to subpar global descriptors. In this paper, we propose a group convolutional approach to multiple view aggregation where convolutions are performed over a discrete subgroup of the rotation group, enabling, thus, joint reasoning over all views in an equivariant (instead of invariant) fashion, up to the very last layer. We further develop this idea to operate on smaller discrete homogeneous spaces of the rotation group, where a polar view representation is used to maintain equivariance with only a fraction of the number of input views. We set the new state of the art in several large scale 3D shape retrieval tasks, and show additional applications to panoramic scene classification.

****

Interpolated Convolutional Networks for 3D Point Cloud Understanding

Jiageng Mao, Xiaogang Wang, Hongsheng Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1578-1587

Point cloud is an important type of 3D representation. However, directly applying convolutions on point clouds is challenging due to the sparse, irregular and unordered data structure. In this paper, we propose a novel Interpolated Convolution operation, InterpConv, to tackle the point cloud feature learning and understanding problem. The key idea is to utilize a set of discrete kernel weights and interpolate point features to neighboring kernel-weight coordinates by an interpolation function for convolution. A normalization term is introduced to handle neighborhoods of different sparsity levels. Our InterpConv is shown to be permutation and sparsity invariant, and can directly handle irregular inputs. We further design Interpolated Convolutional Neural Networks (InterpCNNs) based on InterpConv layers to handle point cloud recognition tasks including shape classification, object part segmentation and indoor scene semantic parsing. Experiments show that the networks can capture both fine-grained local structures and global shape context information effectively. The proposed approach achieves state-of-the-art performance on public benchmarks including ModelNet40, ShapeNet Parts and S3DIS.

****

Revisiting Point Cloud Classification: A New Benchmark Dataset and Classification Model on Real-World Data

Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, Sai-Kit Yeung; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1588-1597

Deep learning techniques for point cloud data have demonstrated great potentials in solving classical problems in 3D computer vision such as 3D object classification and segmentation. Several recent 3D object classification methods have reported state-of-the-art performance on CAD model datasets such as ModelNet40 with

high accuracy (92%). Despite such impressive results, in this paper, we argue that object classification is still a challenging task when objects are framed with real-world settings. To prove this, we introduce ScanObjectNN, a new real-world point cloud object dataset based on scanned indoor scene data. From our comprehensive benchmark, we show that our dataset poses great challenges to existing point cloud classification techniques as objects from real-world scans are often cluttered with background and/or are partial due to occlusions. We identify three key open problems for point cloud object classification, and propose new point cloud classification neural networks that achieve state-of-the-art performance on classifying objects with cluttered background. Our dataset and code are publicly available in our project page https://hkust-vgd.github.io/scanobjectnn/.

****

PointCloud Saliency Maps

Tianhang Zheng, Changyou Chen, Junsong Yuan, Bo Li, Kui Ren; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1598-1606

3D point-cloud recognition with PointNet and its variants has received remarkable progress. A missing ingredient, however, is the ability to automatically evaluate point-wise importance w.r.t. classification performance, which is usually reflected by a saliency map. A saliency map is an important tool as it allows one to perform further processes on point-cloud data. In this paper, we propose a novel way of characterizing critical points and segments to build point-cloud saliency maps. Our method assigns each point a score reflecting its contribution to the model-recognition loss. The saliency map explicitly explains which points are the key for model recognition. Furthermore, aggregations of highly-scored points indicate important segments/subsets in a point-cloud. Our motivation for constructing a saliency map is by point dropping, which is a non-differentiable operator. To overcome this issue, we approximate point-dropping with a differentiable procedure of shifting points towards the cloud centroid. Consequently, each saliency score can be efficiently measured by the corresponding gradient of the loss w.r.t the point under the spherical coordinates. Extensive evaluations on several state-of-the-art point-cloud recognition models, including PointNet, PointNet++ and DGCNN, demonstrate the veracity and generality of our proposed saliency map. Code for experiments is released on https://github.com/tianzheng4/PointCloud-Saliency-Maps

****

ShellNet: Efficient Point Cloud Convolutional Neural Networks Using Concentric Shells Statistics

Zhiyuan Zhang, Binh-Son Hua, Sai-Kit Yeung; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1607-1616

Deep learning with 3D data has progressed significantly since the introduction of convolutional neural networks that can handle point order ambiguity in point cloud data. While being able to achieve good accuracies in various scene understanding tasks, previous methods often have low training speed and complex network architecture. In this paper, we address these problems by proposing an efficient end-to-end permutation invariant convolution for point cloud deep learning. Our simple yet effective convolution operator named ShellConv uses statistics from concentric spherical shells to define representative features and resolve the point order ambiguity, allowing traditional convolution to perform on such features. Based on ShellConv we further build an efficient neural network named ShellNet to directly consume the point clouds with larger receptive fields while maintaining less layers. We demonstrate the efficacy of ShellNet by producing state-of-the-art results on object classification, object part segmentation, and semantic scene segmentation while keeping the network very fast to train.

****

Unsupervised Deep Learning for Structured Shape Matching

Jean-Michel Roufosse, Abhishek Sharma, Maks Ovsjanikov; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1617-1627

We present a novel method for computing correspondences across 3D shapes using unsupervised learning. Our method computes a non-linear transformation of given d

escriptor functions, while optimizing for global structural properties of the re
sulting maps, such as their bijectivity or approximate isometry. To this end, we
 use the functional maps framework, and build upon the recent FMNet architecture
 for descriptor learning. Unlike that approach, however, we show that learning c
an be done in a purely unsupervised setting, without having access to any ground
 truth correspondences. This results in a very general shape matching method tha
t we call SURFMNet for Spectral Unsupervised FMNet, and which can be used to est
ablish correspondences within 3D shape collections without any prior information
. We demonstrate on a wide range of challenging benchmarks, that our approach le
ads to state-of-the-art results compared to the existing unsupervised methods an
d achieves results that are comparable even to the supervised learning technique
s. Moreover, our framework is an order of magnitude faster, and does not rely on
 geodesic distance computation or expensive post-processing.
****

Linearly Converging Quasi Branch and Bound Algorithms for Global Rigid Registrat
ion

Nadav Dym, Shahar Ziv Kovalsky; Proceedings of the IEEE/CVF International Confe
rence on Computer Vision (ICCV), 2019, pp. 1628-1636

In recent years, several branch-and-bound (BnB) algorithms have been proposed to
 globally optimize rigid registration problems. In this paper, we suggest a gene
ral framework to improve upon the BnB approach, which we name Quasi BnB. Quasi B
nB replaces the linear lower bounds used in BnB algorithms with quadratic quasi-
lower bounds which are based on the quadratic behavior of the energy in the vici
nity of the global minimum. While quasi-lower bounds are not truly lower bounds,
 the Quasi-BnB algorithm is globally optimal. In fact we prove that it exhibits
linear convergence -- it achieves epsilon accuracy in O(log(1/epsilon)) time whi
le the time complexity of other rigid registration BnB algorithms is polynomial
in 1/epsilon. Our experiments verify that Quasi-BnB is significantly more effici
ent than state-of-the-art BnB algorithms, especially for problems where high acc
uracy is desired.
****

Consensus Maximization Tree Search Revisited

Zhipeng Cai, Tat-Jun Chin, Vladlen Koltun; Proceedings of the IEEE/CVF Interna
tional Conference on Computer Vision (ICCV), 2019, pp. 1637-1645

Consensus maximization is widely used for robust fitting in computer vision. How
ever, solving it exactly, i.e., finding the globally optimal solution, is intrac
table. A* tree search, which has been shown to be fixed-parameter tractable, is
one of the most efficient exact methods, though it is still limited to small inp
uts. We make two key contributions towards improving A* tree search. First, we s
how that the consensus maximization tree structure used previously actually cont
ains paths that connect nodes at both adjacent and non-adjacent levels. Cruciall
y, paths connecting non-adjacent levels are redundant for tree search, but they
were not avoided previously. We propose a new acceleration strategy that avoids
such redundant paths. In the second contribution, we show that the existing bran
ch pruning technique also deteriorates quickly with the problem dimension. We th
en propose a new branch pruning technique that is less dimension-sensitive to ad
dress this issue. Experiments show that both new techniques can significantly ac
celerate A* tree search, making it reasonably efficient on inputs that were prev
iously out of reach. Demo code is available at https://github.com/ZhipengCai/Max
ConTreeSearch.
****

Quasi-Globally Optimal and Efficient Vanishing Point Estimation in Manhattan Wor
ld

Haoang Li, Ji Zhao, Jean-Charles Bazin, Wen Chen, Zhe Liu, Yun-Hui Liu; Pro
ceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 201
9, pp. 1646-1654

The image lines projected from parallel 3D lines intersect at a common point cal
led the vanishing point (VP). Manhattan world holds for the scenes with three or
thogonal VPs. In Manhattan world, given several lines in a calibrated image, we
aim at clustering them by three unknown-but-sought VPs. The VP estimation can be

reformulated as computing the rotation between the Manhattan frame and the camera frame. To compute this rotation, state-of-the-art methods are based on either data sampling or parameter search, and they fail to guarantee the accuracy and efficiency simultaneously. In contrast, we propose to hybridize these two strategies. We first compute two degrees of freedom (DOF) of the above rotation by two sampled image lines, and then search for the optimal third DOF based on the branch-and-bound. Our sampling accelerates our search by reducing the search space and simplifying the bound computation. Our search is not sensitive to noise and achieves quasi-global optimality in terms of maximizing the number of inliers. Experiments on synthetic and real-world images showed that our method outperforms state-of-the-art approaches in terms of accuracy and/or efficiency.

****

An Efficient Solution to the Homography-Based Relative Pose Problem With a Common Reference Direction

Yaqing Ding, Jian Yang, Jean Ponce, Hui Kong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1655-1664

In this paper, we propose a novel approach to two-view minimal-case relative pose problems based on homography with a common reference direction. We explore the rank-1 constraint on the difference between the Euclidean homography matrix and the corresponding rotation, and propose an efficient two-step solution for solving both the calibrated and partially calibrated (unknown focal length) problems. We derive new 3.5-point, 3.5-point, 4-point solvers for two cameras such that the two focal lengths are unknown but equal, one of them is unknown, and both are unknown and possibly different, respectively. We present detailed analyses and comparisons with existing 6 and 7-point solvers, including results with smart phone images.

****

A Quaternion-Based Certifiably Optimal Solution to the Wahba Problem With Outliers

Heng Yang, Luca Carlone; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1665-1674

The Wahba problem, also known as rotation search, seeks to find the best rotation to align two sets of vector observations given putative correspondences, and is a fundamental routine in many computer vision and robotics applications. This work proposes the first polynomial-time certifiably optimal approach for solving the Wahba problem when a large number of vector observations are outliers. Our first contribution is to formulate the Wahba problem using a Truncated Least Squares (TLS) cost that is insensitive to a large fraction of spurious correspondences. The second contribution is to rewrite the problem using unit quaternions and show that the TLS cost can be framed as a Quadratically-Constrained Quadratic Program (QCQP). Since the resulting optimization is still highly non-convex and hard to solve globally, our third contribution is to develop a convex Semidefinite Programming (SDP) relaxation. We show that while a naive relaxation performs poorly in general, our relaxation is tight even in the presence of large noise and outliers. We validate the proposed algorithm, named QUASAR (QUAternion-based Semidefinite relAxation for Robust alignment), in both synthetic and real datasets showing that the algorithm outperforms RANSAC, robust local optimization techniques, global outlier-removal procedures, and Branch-and-Bound methods. QUASAR is able to compute certifiably optimal solutions (i.e. the relaxation is exact) even in the case when 95% of the correspondences are outliers.

****

PLMP - Point-Line Minimal Problems in Complete Multi-View Visibility

Timothy Duff, Kathlen Kohn, Anton Leykin, Tomas Pajdla; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1675-1684

We present a complete classification of all minimal problems for generic arrangements of points and lines completely observed by calibrated perspective cameras. We show that there are only 30 minimal problems in total, no problems exist for more than 6 cameras, for more than 5 points, and for more than 6 lines. We present a sequence of tests for detecting minimality starting with counting degrees of freedom and ending with full symbolic and numeric verification of representat

ive examples. For all minimal problems discovered, we present their algebraic de grees, i.e. the number of solutions, which measure their intrinsic difficulty. I t shows how exactly the difficulty of problems grows with the number of views. I mportantly, several new mini- mal problems have small degrees that might be prac tical in image matching and 3D reconstruction.
****

Variational Few-Shot Learning
Jian Zhang, Chenglong Zhao, Bingbing Ni, Minghao Xu, Xiaokang Yang; Proceedi ngs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp . 1685-1694
We propose a variational Bayesian framework for enhancing few-shot learning perf ormance. This idea is motivated by the fact that single point based metric learn ing approaches are inherently noise-vulnerable and easy-to-be-biased. In a nutsh ell, stochastic variational inference is invoked to approximate bias-eliminated class specific sample distributions. In the meantime, a classifier-free predicti on is attained by leveraging the distribution statistics on novel samples. Exten sive experimental results on several benchmarks well demonstrate the effectivene ss of our distribution-driven few-shot learning framework over previous point es timates based methods, in terms of superior classification accuracy and robustne ss.
****

Generative Adversarial Minority Oversampling
Sankha Subhra Mullick, Shounak Datta, Swagatam Das; Proceedings of the IEEE/CV F International Conference on Computer Vision (ICCV), 2019, pp. 1695-1704
Class imbalance is a long-standing problem relevant to a number of real-world ap plications of deep learning. Oversampling techniques, which are effective for ha ndling class imbalance in classical learning systems, can not be directly applie d to end-to-end deep learning systems. We propose a three-player adversarial gam e between a convex generator, a multi-class classifier network, and a real/fake discriminator to perform oversampling in deep learning systems. The convex gener ator generates new samples from the minority classes as convex combinations of e xisting instances, aiming to fool both the discriminator as well as the classifi er into misclassifying the generated samples. Consequently, the artificial sampl es are generated at critical locations near the peripheries of the classes. This , in turn, adjusts the classifier induced boundaries in a way which is more like ly to reduce misclassification from the minority classes. Extensive experiments on multiple class imbalanced image datasets establish the efficacy of our propos al.
****

Memorizing Normality to Detect Anomaly: Memory-Augmented Deep Autoencoder for Un supervised Anomaly Detection
Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Sv etha Venkatesh, Anton van den Hengel; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1705-1714
Deep autoencoder has been extensively used for anomaly detection. Training on th e normal data, the autoencoder is expected to produce higher reconstruction erro r for the abnormal inputs than the normal ones, which is adopted as a criterion for identifying anomalies. However, this assumption does not always hold in prac tice. It has been observed that sometimes the autoencoder "generalizes" so well that it can also reconstruct anomalies well, leading to the miss detection of an omalies. To mitigate this drawback for autoencoder based anomaly detector, we pr opose to augment the autoencoder with a memory module and develop an improved au toencoder called memory-augmented autoencoder, i.e. MemAE. Given an input, MemAE firstly obtains the encoding from the encoder and then uses it as a query to re trieve the most relevant memory items for reconstruction. At the training stage, the memory contents are updated and are encouraged to represent the prototypica l elements of the normal data. At the test stage, the learned memory will be fix ed, and the reconstruction is obtained from a few selected memory records of the normal data. The reconstruction will thus tend to be close to a normal sample. Thus the reconstructed errors on anomalies will be strengthened for anomaly dete

ction. MemAE is free of assumptions on the data type and thus general to be appl ied to different tasks. Experiments on various datasets prove the excellent gene ralization and high effectiveness of the proposed MemAE.
****

Topological Map Extraction From Overhead Images
Zuoyue Li, Jan Dirk Wegner, Aurelien Lucchi; Proceedings of the IEEE/CVF Inter national Conference on Computer Vision (ICCV), 2019, pp. 1715-1724
We propose a new approach, named PolyMapper, to circumvent the conventional pixe l-wise segmentation of (aerial) images and predict objects in a vector represent ation directly. PolyMapper directly extracts the topological map of a city from overhead images as collections of building footprints and road networks. In orde r to unify the shape representation for different types of objects, we also prop ose a novel sequentialization method that reformulates a graph structure as clos ed polygons. Experiments are conducted on both existing and self-collected large -scale datasets of several cities. Our empirical results demonstrate that our en d-to-end learnable model is capable of drawing polygons of building footprints a nd road networks that very closely approximate the structure of existing online map services, in a fully automated manner. Quantitative and qualitative comparis on to the state-of-the-arts also show that our approach achieves good levels of performance. To the best of our knowledge, the automatic extraction of large-sca le topological maps is a novel contribution in the remote sensing community that we believe will help develop models with more informed geometrical constraints.
****

Exploiting Temporal Consistency for Real-Time Video Depth Estimation
Haokui Zhang, Chunhua Shen, Ying Li, Yuanzhouhan Cao, Yu Liu, Youliang Yan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1725-1734
Accuracy of depth estimation from static images has been significantly improved recently, by exploiting hierarchical features from deep convolutional neural net works (CNNs). Compared with static images, vast information exists among video f rames and can be exploited to improve the depth estimation performance. In this work, we focus on exploring temporal information from monocular videos for depth estimation. Specifically, we take the advantage of convolutional long short-ter m memory (CLSTM) and propose a novel spatial-temporal CSLTM (ST-CLSTM) structure . Our ST-CLSTM structure can capture not only the spatial features but also the temporal correlations/consistency among consecutive video frames with negligible increase in computational cost. Additionally, in order to maintain the temporal consistency among the estimated depth frames, we apply the generative adversari al learning scheme and design a temporal consistency loss. The temporal consiste ncy loss is combined with the spatial loss to update the model in an end-to-end fashion. By taking advantage of the temporal information, we build a video depth estimation framework that runs in real-time and generates visually pleasant res ults. Moreover, our approach is flexible and can be generalized to most existing depth estimation frameworks. Code is available at: https://tinyurl.com/STCLSTM
****

The Sound of Motions
Hang Zhao, Chuang Gan, Wei-Chiu Ma, Antonio Torralba; Proceedings of the IEEE /CVF International Conference on Computer Vision (ICCV), 2019, pp. 1735-1744
Sounds originate from object motions and vibrations of surrounding air. Inspired by the fact that humans is capable of interpreting sound sources from how objec ts move visually, we propose a novel system that explicitly captures such motion cues for the task of sound localization and separation. Our system is composed of an end-to-end learnable model called Deep Dense Trajectory (DDT), and a curri culum learning scheme. It exploits the inherent coherence of audio-visual signal s from a large quantities of unlabeled videos. Quantitative and qualitative eval uations show that comparing to previous models that rely on visual appearance cu es, our motion based system improves performance in separating musical instrumen t sounds. Furthermore, it separates sound components from duets of the same cate gory of instruments, a challenging problem that has not been addressed before.
****

SC-FEGAN: Face Editing Generative Adversarial Network With User's Sketch and Color

Youngjoo Jo, Jongyoul Park; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1745-1753

We present a novel image editing system that generates images as the user provides free-form masks, sketches and color as inputs. Our system consists of an end-to-end trainable convolutional network. In contrast to the existing methods, our system utilizes entirely free-form user input in terms of color and shape. This allows the system to respond to the user's sketch and color inputs, using them as guidelines to generate an image. In this work, we trained the network with an additional style loss, which made it possible to generate realistic results despite large portions of the image being removed. Our proposed network architecture SC-FEGAN is well suited for generating high-quality synthetic images using intuitive user inputs.

****

Exploring Overall Contextual Information for Image Captioning in Human-Like Cognitive Style

Hongwei Ge, Zehang Yan, Kai Zhang, Mingde Zhao, Liang Sun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1754-1763

Image captioning is a research hotspot where encoder-decoder models combining convolutional neural network (CNN) and long short-term memory (LSTM) achieve promising results. Despite significant progress, these models generate sentences differently from human cognitive styles. Existing models often generate a complete sentence from the first word to the end, without considering the influence of the following words on the whole sentence generation. In this paper, we explore the utilization of a human-like cognitive style, i.e., building overall cognition for the image to be described and the sentence to be constructed, for enhancing computer image understanding. This paper first proposes a Mutual-aid network structure with Bidirectional LSTMs (MaBi-LSTMs) for acquiring overall contextual information. In the training process, the forward and backward LSTMs encode the succeeding and preceding words into their respective hidden states by simultaneously constructing the whole sentence in a complementary manner. In the captioning process, the LSTM implicitly utilizes the subsequent semantic information contained in its hidden states. In fact, MaBi-LSTMs can generate two sentences in forward and backward directions. To bridge the gap between cross-domain models and generate a sentence with higher quality, we further develop a cross-modal attention mechanism to retouch the two sentences by fusing their salient parts as well as the salient areas of the image. Experimental results on the Microsoft COCO dataset show that the proposed model improves the performance of encoder-decoder models and achieves state-of-the-art results.

****

Order-Aware Generative Modeling Using the 3D-Craft Dataset

Zhuoyuan Chen, Demi Guo, Tong Xiao, Saining Xie, Xinlei Chen, Haonan Yu, Jonathan Gray, Kavya Srinet, Haoqi Fan, Jerry Ma, Charles R. Qi, Shubham Tulsiani, Arthur Szlam, C. Lawrence Zitnick; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1764-1773

In this paper, we study the problem of sequentially building houses in the game of Minecraft, and demonstrate that learning the ordering can make for more effective autoregressive models. Given a partially built house made by a human player, our system tries to place additional blocks in a human-like manner to complete the house. We introduce a new dataset, HouseCraft, for this new task. HouseCraft contains the sequential order in which 2,500 Minecraft houses were built from scratch by humans. The human action sequences enable us to learn an order-aware generative model called Voxel-CNN. In contrast to many generative models where the sequential generation ordering either does not matter (e.g. holistic generation with GANs), or is manually/arbitrarily set by simple rules (e.g. raster-scan order), our focus is on an ordered generation that imitates humans. To evaluate if a generative model can accurately predict human-like actions, we propose several novel quantitative metrics. We demonstrate that our Voxel-CNN model is simpl

e and effective at this creative task, and can serve as a strong baseline for fu ture research in this direction. The HouseCraft dataset and code with baseline m odels will be made publicly available.
****

Crowd Counting With Deep Structured Scale Integration Network
Lingbo Liu, Zhilin Qiu, Guanbin Li, Shufan Liu, Wanli Ouyang, Liang Lin; Pr oceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 20 19, pp. 1774-1783
Automatic estimation of the number of people in unconstrained crowded scenes is a challenging task and one major difficulty stems from the huge scale variation of people. In this paper, we propose a novel Deep Structured Scale Integration N etwork (DSSINet) for crowd counting, which addresses the scale variation of peop le by using structured feature representation learning and hierarchically struct ured loss function optimization. Unlike conventional methods which directly fuse multiple features with weighted average or concatenation, we first introduce a Structured Feature Enhancement Module based on conditional random fields (CRFs) to refine multiscale features mutually with a message passing mechanism. Specifi cally, each scale-specific feature is considered as a continuous random variable and passes complementary information to refine the features at other scales. Se cond, we utilize a Dilated Multiscale Structural Similarity loss to enforce our DSSINet to learn the local correlation of people's scales within regions of vari ous size, thus yielding high-quality density maps. Extensive experiments on four challenging benchmarks well demonstrate the effectiveness of our method. In par ticular, our DSSINet achieves improvements of 9.5% error reduction on Shanghaite ch dataset and 24.9% on UCF-QNRF dataset against the state-of-the-art methods.
****

Bidirectional One-Shot Unsupervised Domain Mapping
Tomer Cohen, Lior Wolf; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1784-1792
We study the problem of mapping between a domain A, in which there is a single t raining sample and a domain B, for which we have a richer training set. The meth od we present is able to perform this mapping in both directions. For example, w e can transfer all MNIST images to the visual domain captured by a single SVHN i mage and transform the SVHN image to the domain of the MNIST images. Our method is based on employing one encoder and one decoder for each domain, without utili zing weight sharing. The autoencoder of the single sample domain is trained to m atch both this sample and the latent space of domain B. Our results demonstrate convincing mapping between domains, where either the source or the target domain are defined by a single sample, far surpassing existing solutions. Our code is made publicly available at https://github.com/tomercohen11/BiOST.
****

Evolving Space-Time Neural Architectures for Videos
AJ Piergiovanni, Anelia Angelova, Alexander Toshev, Michael S. Ryoo; Proceedi ngs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp . 1793-1802
We present a new method for finding video CNN architectures that more optimally capture rich spatio-temporal information in videos. Previous work, taking advant age of 3D convolutions, obtained promising results by manually designing CNN vid eo architectures. We here develop a novel evolutionary algorithm that automatica lly explores models with different types and combinations of layers to jointly l earn interactions between spatial and temporal aspects of video representations. We demonstrate the generality of this algorithm by applying it to two meta-arch itectures. Further, we propose a new component, the iTGM layer, which more effic iently utilizes its parameters to allow learning of space-time interactions over longer time horizons. The iTGM layer is often preferred by the evolutionary alg orithm and allows building cost-efficient networks. The proposed approach discov ers new diverse and interesting video architectures that were unknown previously . More importantly they are both more accurate and faster than prior models, and outperform the state-of-the-art results on four datasets: Kinetics, Charades, M oments in Time and HMDB. We will open source the code and models, to encourage f

uture model development.
****
Universally Slimmable Networks and Improved Training Techniques
Jiahui Yu, Thomas S. Huang; Proceedings of the IEEE/CVF International Conferenc e on Computer Vision (ICCV), 2019, pp. 1803-1811
Slimmable networks are a family of neural networks that can instantly adjust the runtime width. The width can be chosen from a predefined widths set to adaptive ly optimize accuracy-efficiency trade-offs at runtime. In this work, we propose a systematic approach to train universally slimmable networks (US-Nets), extendi ng slimmable networks to execute at arbitrary width, and generalizing to network s both with and without batch normalization layers. We further propose two impro ved training techniques for US-Nets, named the sandwich rule and inplace distill ation, to enhance training process and boost testing accuracy. We show improved performance of universally slimmable MobileNet v1 and MobileNet v2 on ImageNet c lassification task, compared with individually trained ones and 4-switch slimmab le network baselines. We also evaluate the proposed US-Nets and improved trainin g techniques on tasks of image super-resolution and deep reinforcement learning. Extensive ablation experiments on these representative tasks demonstrate the ef fectiveness of our proposed methods. Our discovery opens up the possibility to d irectly evaluate FLOPs-Accuracy spectrum of network architectures. Code and mode ls are available at: https://github.com/JiahuiYu/slimmable_networks.
****
AutoDispNet: Improving Disparity Estimation With AutoML
Tonmoy Saikia, Yassine Marrakchi, Arber Zela, Frank Hutter, Thomas Brox; Pro ceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 201 9, pp. 1812-1823
Much research work in computer vision is being spent on optimizing existing netw ork architectures to obtain a few more percentage points on benchmarks. Recent A utoML approaches promise to relieve us from this effort. However, they are mainl y designed for comparatively small-scale classification tasks. In this work, we show how to use and extend existing AutoML techniques to efficiently optimize la rge-scale U-Net-like encoder-decoder architectures. In particular, we leverage g radient-based neural architecture search and Bayesian optimization for hyperpara meter search. The resulting optimization does not require a large-scale compute cluster. We show results on disparity estimation that clearly outperform the man ually optimized baseline and reach state-of-the-art performance.
****
Deep Meta Functionals for Shape Representation
Gidi Littwin, Lior Wolf; Proceedings of the IEEE/CVF International Conference o n Computer Vision (ICCV), 2019, pp. 1824-1833
We present a new method for 3D shape reconstruction from a single image, in whic h a deep neural network directly maps an image to a vector of network weights. T he network parametrized by these weights represents a 3D shape by classifying ev ery point in the volume as either within or outside the shape. The new represent ation has virtually unlimited capacity and resolution, and can have an arbitrary topology. Our experiments show that it leads to more accurate shape inference f rom a 2D projection than the existing methods, including voxel-, silhouette-, an d mesh-based methods. The code will be available at: https: //github.com/gidilit twin/Deep-Meta.
****
Differentiable Kernel Evolution
Yu Liu, Jihao Liu, Ailing Zeng, Xiaogang Wang; Proceedings of the IEEE/CVF In ternational Conference on Computer Vision (ICCV), 2019, pp. 1834-1843
This paper proposes a differentiable kernel evolution (DKE) algorithm to find a better layer-operator for the convolutional neural network. Unlike most of the o ther neural architecture searching (NAS) technologies, we consider the searching space in a fundamental scope: kernel space, which encodes the assembly of basic multiply-accumulate (MAC) operations into a conv-kernel. We first deduce a stri ct form of the generalized convolutional operator by some necessary constraints and construct a continuous searching space for its extra freedom-of-degree, name

ly, the connection of each MAC. Then a novel unsupervised greedy evolution algorithm called gradient agreement guided searching (GAGS) is proposed to learn the optimal location for each MAC in the spatially continuous searching space. We leverage DKE on multiple kinds of tasks such as object classification, face/object detection, large-scale fine-grained and recognition, with various kinds of backbone architecture. Not to mention the consistent performance gain, we found the proposed DKE can further act as an auto-dilated operator, which makes it easy to boost the performance of miniaturized neural networks in multiple tasks.
****

Batch Weight for Domain Adaptation With Mass Shift
Mikolaj Binkowski, Devon Hjelm, Aaron Courville; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1844-1853
Unsupervised domain transfer is the task of transferring or translating samples from a source distribution to a different target distribution. Current solutions unsupervised domain transfer often operate on data on which the modes of the distribution are well-matched, for instance have the same frequencies of classes between source and target distributions. However, these models do not perform well when the modes are not well-matched, as would be the case when samples are drawn independently from two different, but related, domains. This mode imbalance is problematic as generative adversarial networks (GANs), a successful approach in this setting, are sensitive to mode frequency, which results in a mismatch of semantics between source samples and generated samples of the target distribution. We propose a principled method of re-weighting training samples to correct for such mass shift between the transferred distributions, which we call batch weight. We also provide rigorous probabilistic setting for domain transfer and new simplified objective for training transfer networks, an alternative to complex, multi-component loss functions used in the current state-of-the art image-to-image translation models. The new objective stems from the discrimination of joint distributions and enforces cycle-consistency in an abstract, high-level, rather than pixel-wise, sense. Lastly, we experimentally show the effectiveness of the proposed methods in several image-to-image translation tasks.
****

SRM: A Style-Based Recalibration Module for Convolutional Neural Networks
HyunJae Lee, Hyo-Eun Kim, Hyeonseob Nam; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1854-1862
Following the advance of style transfer with Convolutional Neural Networks (CNNs), the role of styles in CNNs has drawn growing attention from a broader perspective. In this paper, we aim to fully leverage the potential of styles to improve the performance of CNNs in general vision tasks. We propose a Style-based Recalibration Module (SRM), a simple yet effective architectural unit, which adaptively recalibrates intermediate feature maps by exploiting their styles. SRM first extracts the style information from each channel of the feature maps by style pooling, then estimates per-channel recalibration weight via channel-independent style integration. By incorporating the relative importance of individual styles into feature maps, SRM effectively enhances the representational ability of a CNN. The proposed module is directly fed into existing CNN architectures with negligible overhead. We conduct comprehensive experiments on general image recognition as well as tasks related to styles, which verify the benefit of SRM over recent approaches such as Squeeze-and-Excitation (SE). To explain the inherent difference between SRM and SE, we provide an in-depth comparison of their representational properties.
****

Switchable Whitening for Deep Representation Learning
Xingang Pan, Xiaohang Zhan, Jianping Shi, Xiaoou Tang, Ping Luo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1863-1871
Normalization methods are essential components in convolutional neural networks (CNNs). They either standardize or whiten data using statistics estimated in predefined sets of pixels. Unlike existing works that design normalization techniques for specific tasks, we propose Switchable Whitening (SW), which provides a ge

neral form unifying different whitening methods as well as standardization metho
ds. SW learns to switch among these operations in an end-to-end manner. It has s
everal advantages. First, SW adaptively selects appropriate whitening or standar
dization statistics for different tasks (see Fig.1), making it well suited for a
 wide range of tasks without manual design. Second, by integrating benefits of d
ifferent normalizers, SW shows consistent improvements over its counterparts in
various challenging benchmarks. Third, SW serves as a useful tool for understand
ing the characteristics of whitening and standardization techniques. We show tha
t SW outperforms other alternatives on image classification (CIFAR-10/100, Image
Net), semantic segmentation (ADE20K, Cityscapes), domain adaptation (GTA5, Citys
capes), and image style transfer (COCO). For example, without bells and whistles
, we achieve state-of-the-art performance with 45.33% mIoU on the ADE20K dataset
.
****

Adaptative Inference Cost With Convolutional Neural Mixture Models
Adria Ruiz, Jakob Verbeek; Proceedings of the IEEE/CVF International Conference
 on Computer Vision (ICCV), 2019, pp. 1872-1881
Despite the outstanding performance of convolutional neural networks (CNNs) for
many vision tasks, the required computational cost during inference is problemat
ic when resources are limited. In this context, we propose Convolutional Neural
Mixture Models (CNMMs), a probabilistic model embedding a large number of CNNs t
hat can be jointly trained and evaluated in an efficient manner. Within the prop
osed framework, we present different mechanisms to prune subsets of CNNs from th
e mixture, allowing to easily adapt the computational cost required for inferenc
e. Image classification and semantic segmentation experiments show that our meth
od achieve excellent accuracy-compute trade-offs. Moreover, unlike most of previ
ous approaches, a single CNMM provides a large range of operating points along t
his trade-off, without any re-training.
****

On Network Design Spaces for Visual Recognition
Ilija Radosavovic, Justin Johnson, Saining Xie, Wan-Yen Lo, Piotr Dollar; Pr
oceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 20
19, pp. 1882-1890
Over the past several years progress in designing better neural network architec
tures for visual recognition has been substantial. To help sustain this rate of
progress, in this work we propose to reexamine the methodology for comparing net
work architectures. In particular, we introduce a new comparison paradigm of dis
tribution estimates, in which network design spaces are compared by applying sta
tistical techniques to populations of sampled models, while controlling for conf
ounding factors like network complexity. Compared to current methodologies of co
mparing point and curve estimates of model families, distribution estimates pain
t a more complete picture of the entire design landscape. As a case study, we ex
amine design spaces used in neural architecture search (NAS). We find significan
t statistical differences between recent NAS design space variants that have bee
n largely overlooked. Furthermore, our analysis reveals that the design spaces f
or standard model families like ResNeXt can be comparable to the more complex on
es used in recent NAS work. We hope these insights into distribution analysis wi
ll enable more robust progress toward discovering better networks for visual rec
ognition.
****

Improved Techniques for Training Adaptive Deep Networks
Hao Li, Hong Zhang, Xiaojuan Qi, Ruigang Yang, Gao Huang; Proceedings of the
 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1891-190
0
Adaptive inference is a promising technique to improve the computational efficie
ncy of deep models at test time. In contrast to static models which use the same
 computation graph for all instances, adaptive networks can dynamically adjust t
heir structure conditioned on each input. While existing research on adaptive in
ference mainly focuses on designing more advanced architectures, this paper inve
stigates how to train such networks more effectively. Specifically, we consider

a typical adaptive deep network with multiple intermediate classifiers. We prese nt three techniques to improve its training efficacy from two aspects: 1) a Grad ient Equilibrium algorithm to resolve the conflict of learning of different clas sifiers; 2) an Inline Subnetwork Collaboration approach and a One-for-all Knowle dge Distillation algorithm to enhance the collaboration among classifiers. On mu ltiple datasets (CIFAR-10, CIFAR-100 and ImageNet), we show that the proposed ap proach consistently leads to further improved efficiency on top of state-of-the- art adaptive deep networks.

****

Resource Constrained Neural Network Architecture Search: Will a Submodularity As sumption Help?

Yunyang Xiong, Ronak Mehta, Vikas Singh; Proceedings of the IEEE/CVF Internati onal Conference on Computer Vision (ICCV), 2019, pp. 1901-1910

The design of neural network architectures is frequently either based on human e xpertise using trial/error and empirical feedback or tackled via large scale rei nforcement learning strategies performed over distinct discrete architecture cho ices. In the latter case, the optimization is often non-differentiable and also not very amenable to derivative-free optimization methods. Most methods in use t oday require sizable computational resources. And if we want networks that addit ionally satisfy resource constraints, the above challenges are exacerbated becau se the search must now balance accuracy with certain budget constraints on resou rces. We formulate this problem as the optimization of a set function -- we find that the empirical behavior of this set function often (but not always) satisfi es marginal gain and monotonicity principles -- properties central to the idea o f submodularity. Based on this observation, we adapt algorithms within discrete optimization to obtain heuristic schemes for neural network architecture search, where we have resource constraints on the architecture. This simple scheme when applied on CIFAR-100 and ImageNet, identifies resource-constrained architecture s with quantifiably better performance than current state-of-the-art models desi gned for mobile devices. Specifically, we find high-performing architectures wit h fewer parameters and computations by a search method that is much faster.

****

ACNet: Strengthening the Kernel Skeletons for Powerful CNN via Asymmetric Convol ution Blocks

Xiaohan Ding, Yuchen Guo, Guiguang Ding, Jungong Han; Proceedings of the IEEE /CVF International Conference on Computer Vision (ICCV), 2019, pp. 1911-1920

As designing appropriate Convolutional Neural Network (CNN) architecture in the context of a given application usually involves heavy human works or numerous GP U hours, the research community is soliciting the architecture-neutral CNN struc tures, which can be easily plugged into multiple mature architectures to improve the performance on our real-world applications. We propose Asymmetric Convoluti on Block (ACB), an architecture-neutral structure as a CNN building block, which uses 1D asymmetric convolutions to strengthen the square convolution kernels. F or an off-the-shelf architecture, we replace the standard square-kernel convolut ional layers with ACBs to construct an Asymmetric Convolutional Network (ACNet), which can be trained to reach a higher level of accuracy. After training, we eq uivalently convert the ACNet into the same original architecture, thus requiring no extra computations anymore. We have observed that ACNet can improve the perf ormance of various models on CIFAR and ImageNet by a clear margin. Through furth er experiments, we attribute the effectiveness of ACB to its capability of enhan cing the model's robustness to rotational distortions and strengthening the cent ral skeleton parts of square convolution kernels.

****

A Comprehensive Overhaul of Feature Distillation

Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, Jin Young C hoi; Proceedings of the IEEE/CVF International Conference on Computer Vision (IC CV), 2019, pp. 1921-1930

We investigate the design aspects of feature distillation methods achieving netw ork compression and propose a novel feature distillation method in which the dis tillation loss is designed to make a synergy among various aspects: teacher tran

sform, student transform, distillation feature position and distance function. Our proposed distillation loss includes a feature transform with a newly designed margin ReLU, a new distillation feature position, and a partial L2 distance function to skip redundant information giving adverse effects to the compression of student. In ImageNet, our proposed method achieves 21.65% of top-1 error with ResNet50, which outperforms the performance of the teacher network, ResNet152. Our proposed method is evaluated on various tasks such as image classification, object detection and semantic segmentation and achieves a significant performance improvement in all tasks. The code is available at project page.
****

Transferable Semi-Supervised 3D Object Detection From RGB-D Data
Yew Siang Tang, Gim Hee Lee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1931-1940
We investigate the direction of training a 3D object detector for new object classes from only 2D bounding box labels of these new classes, while simultaneously transferring information from 3D bounding box labels of the existing classes. To this end, we propose a transferable semi-supervised 3D object detection model that learns a 3D object detector network from training data with two disjoint sets of object classes - a set of strong classes with both 2D and 3D box labels, and another set of weak classes with only 2D box labels. In particular, we suggest a relaxed reprojection loss, box prior loss and a Box-to-Point Cloud Fit network that allow us to effectively transfer useful 3D information from the strong classes to the weak classes during training, and consequently, enable the network to detect 3D objects in the weak classes during inference. Experimental results show that our proposed algorithm outperforms baseline approaches and achieves promising results compared to fully-supervised approaches on the SUN-RGBD and KITTI datasets. Furthermore, we show that our Box-to-Point Cloud Fit network improves performances of the fully-supervised approaches on both datasets.
****

DPOD: 6D Pose Object Detector and Refiner
Sergey Zakharov, Ivan Shugurov, Slobodan Ilic; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1941-1950
In this paper we present a novel deep learning method for 3D object detection and 6D pose estimation from RGB images. Our method, named DPOD (Dense Pose Object Detector), estimates dense multi-class 2D-3D correspondence maps between an input image and available 3D models. Given the correspondences, a 6DoF pose is computed via PnP and RANSAC. An additional RGB pose refinement of the initial pose estimates is performed using a custom deep learning-based refinement scheme. Our results and comparison to a vast number of related works demonstrate that a large number of correspondences is beneficial for obtaining high-quality 6D poses both before and after refinement. Unlike other methods that mainly use real data for training and do not train on synthetic renderings, we perform evaluation on both synthetic and real training data demonstrating superior results before and after refinement when compared to all recent detectors. While being precise, the presented approach is still real-time capable.
****

STD: Sparse-to-Dense 3D Object Detector for Point Cloud
Zetong Yang, Yanan Sun, Shu Liu, Xiaoyong Shen, Jiaya Jia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1951-1960
We propose a two-stage 3D object detection framework, named sparse-to-dense 3D Object Detector (STD). The first stage is a bottom-up proposal generation network that uses raw point clouds as input to generate accurate proposals by seeding each point with a new spherical anchor. It achieves a higher recall with less computation compared with prior works. Then, PointsPool is applied for proposal feature generation by transforming interior point features from sparse expression to compact representation, which saves even more computation. In box prediction, which is the second stage, we implement a parallel intersection-over-union (IoU) branch to increase awareness of localization accuracy, resulting in further improved performance. We conduct experiments on KITTI dataset, and evaluate our met

hod on 3D object and Bird's Eye View (BEV) detection. Our method outperforms oth
er methods by a large margin, especially on the hard set, with 10+ FPS inference
 speed.
****

DUP-Net: Denoiser and Upsampler Network for 3D Adversarial Point Clouds Defense
Hang Zhou, Kejiang Chen, Weiming Zhang, Han Fang, Wenbo Zhou, Nenghai Yu; P
roceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2
019, pp. 1961-1970
Neural networks are vulnerable to adversarial examples, which poses a threat to
their application in security sensitive systems. We propose a Denoiser and UPsam
pler Network (DUP-Net) structure as defenses for 3D adversarial point cloud clas
sification, where the two modules reconstruct surface smoothness by dropping or
adding points. In this paper, statistical outlier removal (SOR) and a data-drive
n upsampling network are considered as denoiser and upsampler respectively. Comp
ared with baseline defenses, DUP-Net has three advantages. First, with DUP-Net a
s a defense, the target model is more robust to white-box adversarial attacks. S
econd, the statistical outlier removal provides added robustness since it is a n
on-differentiable denoising operation. Third, the upsampler network can be train
ed on a small dataset and defends well against adversarial attacks generated fro
m other point cloud datasets. We conduct various experiments to validate that DU
P-Net is very effective as defense in practice. Our best defense eliminates 83.8
% of C&W and l2 loss based attack (point shifting), 50.0% of C&W and Hausdorff d
istance loss based attack (point adding) and 9.0% of saliency map based attack (
point dropping) under 200 dropped points on PointNet.
****

Learning Rich Features at High-Speed for Single-Shot Object Detection
Tiancai Wang, Rao Muhammad Anwer, Hisham Cholakkal, Fahad Shahbaz Khan, Yanw
ei Pang, Ling Shao; Proceedings of the IEEE/CVF International Conference on Com
puter Vision (ICCV), 2019, pp. 1971-1980
Single-stage object detection methods have received significant attention recent
ly due to their characteristic realtime capabilities and high detection accuraci
es. Generally, most existing single-stage detectors follow two common practices:
 they employ a network backbone that is pretrained on ImageNet for the classific
ation task and use a top-down feature pyramid representation for handling scale
variations. Contrary to common pre-training strategy, recent works have demonstr
ated the benefits of training from scratch to reduce the task gap between classi
fication and localization, especially at high overlap thresholds. However, detec
tion models trained from scratch require significantly longer training time comp
ared to their typical finetuning based counterparts. We introduce a single-stage
 detection framework that combines the advantages of both fine-tuning pretrained
 models and training from scratch. Our framework constitutes a standard network
that uses a pre-trained backbone and a parallel light-weight auxiliary network t
rained from scratch. Further, we argue that the commonly used top-down pyramid r
epresentation only focuses on passing high-level semantics from the top layers t
o bottom layers. We introduce a bi-directional network that efficiently circulat
es both low-/mid-level and high-level semantic information in the detection fram
ework. Experiments are performed on MS COCO and UAVDT datasets. Compared to the
baseline, our detector achieives an absolute gain of 7.4% and 4.2% in average pr
ecision (AP) on MS COCO and UAVDT datasets, respectively using VGG backbone. For
 a 300x300 input on the MS COCO test set, our detector with ResNet backbone surp
asses existing single-stage detection methods for single-scale inference achievi
ng 34.3 AP, while operating at an inference time of 19 milliseconds on a single
Titan X GPU. Code is avail- able at https://github.com/vaesl/LRF-Net.
****

Detecting Unseen Visual Relations Using Analogies
Julia Peyre, Ivan Laptev, Cordelia Schmid, Josef Sivic; Proceedings of the IE
EE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1981-1990
We seek to detect visual relations in images of the form of triplets t = (subjec
t, predicate, object), such as "person riding dog", where training examples of t
he individual entities are available but their combinations are unseen at traini

ng. This is an important set-up due to the combinatorial nature of visual relations : collecting sufficient training data for all possible triplets would be very hard. The contributions of this work are three-fold. First, we learn a representation of visual relations that combines (i) individual embeddings for subject, object and predicate together with (ii) a visual phrase embedding that represents the relation triplet. Second, we learn how to transfer visual phrase embeddings from existing training triplets to unseen test triplets using analogies between relations that involve similar objects. Third, we demonstrate the benefits of our approach on three challenging datasets : on HICO-DET, our model achieves significant improvement over a strong baseline for both frequent and unseen triplets, and we observe similar improvement for the retrieval of unseen triplets with out-of-vocabulary predicates on the COCO-a dataset as well as the challenging unusual triplets in the UnRel dataset.
****

Disentangling Monocular 3D Object Detection
Andrea Simonelli, Samuel Rota Bulo, Lorenzo Porzi, Manuel Lopez-Antequera, Peter Kontschieder; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1991-1999
In this paper we propose an approach for monocular 3D object detection from a single RGB image, which leverages a novel disentangling transformation for 2D and 3D detection losses and a novel, self-supervised confidence score for 3D bounding boxes. Our proposed loss disentanglement has the twofold advantage of simplifying the training dynamics in the presence of losses with complex interactions of parameters, and sidestepping the issue of balancing independent regression terms. Our solution overcomes these issues by isolating the contribution made by groups of parameters to a given loss, without changing its nature. We further apply loss disentanglement to another novel, signed Intersection-over-Union criterion-driven loss for improving 2D detection results. Besides our methodological innovations, we critically review the AP metric used in KITTI3D, which emerged as the most important dataset for comparing 3D detection results. We identify and resolve a flaw in the 11-point interpolated AP metric, affecting all previously published detection results and particularly biases the results of monocular 3D detection. We provide extensive experimental evaluations and ablation studies and set a new state-of-the-art on the KITTI3D Car class.
****

STM: SpatioTemporal and Motion Encoding for Action Recognition
Boyuan Jiang, MengMeng Wang, Weihao Gan, Wei Wu, Junjie Yan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2000-2009
Spatiotemporal and motion features are two complementary and crucial information for video action recognition. Recent state-of-the-art methods adopt a 3D CNN stream to learn spatiotemporal features and another flow stream to learn motion features. In this work, we aim to efficiently encode these two features in a unified 2D framework. To this end, we first propose a STM block, which contains a Channel-wise SpatioTemporal Module (CSTM) to present the spatiotemporal features and a Channel-wise Motion Module (CMM) to efficiently encode motion features. We then replace original residual blocks in the ResNet architecture with STM blcoks to form a simple yet effective STM network by introducing very limited extra computation cost. Extensive experiments demonstrate that the proposed STM network outperforms the state-of-the-art methods on both temporal-related datasets (i.e., Something-Something v1 & v2 and Jester) and scene-related datasets (i.e., Kinetics-400, UCF-101, and HMDB-51) with the help of encoding spatiotemporal and motion features together.
****

Dynamic Context Correspondence Network for Semantic Alignment
Shuaiyi Huang, Qiuyue Wang, Songyang Zhang, Shipeng Yan, Xuming He; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2010-2019
Establishing semantic correspondence is a core problem in computer vision and remains challenging due to large intra-class variations and lack of annotated data

. In this paper, we aim to incorporate global semantic context in a flexible man ner to overcome the limitations of prior work that relies on local semantic repr esentations. To this end, we first propose a context-aware semantic representati on that incorporates spatial layout for robust matching against local ambiguitie s. We then develop a novel dynamic fusion strategy based on attention mechanism to weave the advantages of both local and context features by integrating semant ic cues from multiple scales. We instantiate our strategy by designing an end-to -end learnable deep network, named as Dynamic Context Correspondence Network (DC CNet). To train the network, we adopt a multi-auxiliary task loss to improve the efficiency of our weakly-supervised learning procedure. Our approach achieves s uperior or competitive performance over previous methods on several challenging datasets, including PF-Pascal, PF-Willow, and TSS, demonstrating its effectivene ss and generality.

****

Fooling Network Interpretation in Image Classification

Akshayvarun Subramanya, Vipin Pillai, Hamed Pirsiavash; Proceedings of the IEE E/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2020-2029

Deep neural networks have been shown to be fooled rather easily using adversaria l attack algorithms. Practical methods such as adversarial patches have been sho wn to be extremely effective in causing misclassification. However, these patche s are highlighted using standard network interpretation algorithms, thus reveali ng the identity of the adversary. We show that it is possible to create adversar ial patches which not only fool the prediction, but also change what we interpre t regarding the cause of the prediction. Moreover, we introduce our attack as a controlled setting to measure the accuracy of interpretation algorithms. We show this using extensive experiments for Grad-CAM interpretation that transfers to occluding patch interpretation as well. We believe our algorithms can facilitate developing more robust network interpretation tools that truly explain the netw ork's underlying decision making process.

****

Unconstrained Foreground Object Search

Yinan Zhao, Brian Price, Scott Cohen, Danna Gurari; Proceedings of the IEEE/C VF International Conference on Computer Vision (ICCV), 2019, pp. 2030-2039

Many people search for foreground objects to use when editing images. While exis ting methods can retrieve candidates to aid in this, they are constrained to ret urning objects that belong to a pre-specified semantic class. We instead propose a novel problem of unconstrained foreground object (UFO) search and introduce a solution that supports efficient search by encoding the background image in the same latent space as the candidate foreground objects. A key contribution of ou r work is a cost-free, scalable approach for creating a large-scale training dat aset with a variety of foreground objects of differing semantic categories per i mage location. Quantitative and human-perception experiments with two diverse da tasets demonstrate the advantage of our UFO search solution over related baselin es.

****

Embodied Amodal Recognition: Learning to Move to Perceive Objects

Jianwei Yang, Zhile Ren, Mingze Xu, Xinlei Chen, David J. Crandall, Devi Pa rikh, Dhruv Batra; Proceedings of the IEEE/CVF International Conference on Comp uter Vision (ICCV), 2019, pp. 2040-2050

Passive visual systems typically fail to recognize objects in the amodal setting where they are heavily occluded. In contrast, humans and other embodied agents have the ability to move in the environment and actively control the viewing ang le to better understand object shapes and semantics. In this work, we introduce the task of Embodied Amodel Recognition (EAR): an agent is instantiated in a 3D environment close to an occluded target object, and is free to move in the envir onment to perform object classification, amodal object localization, and amodal object segmentation. To address this problem, we develop a new model called Embo died Mask R-CNN for agents to learn to move strategically to improve their visua l recognition abilities. We conduct experiments using a simulator for indoor env ironments. Experimental results show that: 1) agents with embodiment (movement)

achieve better visual recognition performance than passive ones and 2) in order to improve visual recognition abilities, agents can learn strategic paths that are different from shortest paths.
****

SpatialSense: An Adversarially Crowdsourced Benchmark for Spatial Relation Recognition
Kaiyu Yang, Olga Russakovsky, Jia Deng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2051-2060
Understanding the spatial relations between objects in images is a surprisingly challenging task. A chair may be "behind" a person even if it appears to the left of the person in the image (depending on which way the person is facing). Two students that appear close to each other in the image may not in fact be "next to" each other if there is a third student between them. We introduce SpatialSense, a dataset specializing in spatial relation recognition which captures a broad spectrum of such challenges, allowing for proper benchmarking of computer vision techniques. SpatialSense is constructed through adversarial crowdsourcing, in which human annotators are tasked with finding spatial relations that are difficult to predict using simple cues such as 2D spatial configuration or language priors. Adversarial crowdsourcing significantly reduces dataset bias and samples more interesting relations in the long tail compared to existing datasets. On SpatialSense, state-of-the-art recognition models perform comparably to simple baselines, suggesting that they rely on straightforward cues instead of fully reasoning about this complex task. The SpatialSense benchmark provides a path forward to advancing the spatial reasoning capabilities of computer vision systems. The dataset and code are available at https://github.com/princeton-vl/SpatialSense.
****

TensorMask: A Foundation for Dense Object Segmentation
Xinlei Chen, Ross Girshick, Kaiming He, Piotr Dollar; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2061-2069
Sliding-window object detectors that generate bounding-box object predictions over a dense, regular grid have advanced rapidly and proven popular. In contrast, modern instance segmentation approaches are dominated by methods that first detect object bounding boxes, and then crop and segment these regions, as popularized by Mask R-CNN. In this work, we investigate the paradigm of dense sliding-window instance segmentation, which is surprisingly under-explored. Our core observation is that this task is fundamentally different than other dense prediction tasks such as semantic segmentation or bounding-box object detection, as the output at every spatial location is itself a geometric structure with its own spatial dimensions. To formalize this, we treat dense instance segmentation as a prediction task over 4D tensors and present a general framework called TensorMask that explicitly captures this geometry and enables novel operators on 4D tensors. We demonstrate that the tensor view leads to large gains over baselines that ignore this structure, and leads to results comparable to Mask R-CNN. These promising results suggest that TensorMask can serve as a foundation for novel advances in dense mask prediction and a more complete understanding of the task. Code will be made available.
****

Integral Object Mining via Online Attention Accumulation
Peng-Tao Jiang, Qibin Hou, Yang Cao, Ming-Ming Cheng, Yunchao Wei, Hong-Kai Xiong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2070-2079
Object attention maps generated by image classifiers are usually used as priors for weakly-supervised segmentation approaches. However, normal image classifiers produce attention only at the most discriminative object parts, which limits the performance of weakly-supervised segmentation task. Therefore, how to effectively identify entire object regions in a weakly-supervised manner has always been a challenging and meaningful problem. We observe that the attention maps produced by a classification network continuously focus on different object parts during training. In order to accumulate the discovered different object parts, we propose an online attention accumulation (OAA) strategy which maintains a cumulati

ve attention map for each target category in each training image so that the int egral object regions can be gradually promoted as the training goes. These cumul ative attention maps, in turn, serve as the pixel-level supervision, which can f urther assist the network in discovering more integral object regions. Our metho d (OAA) can be plugged into any classification network and progressively accumul ate the discriminative regions into integral objects as the training process goe s. Despite its simplicity, when applying the resulting attention maps to the wea kly-supervised semantic segmentation task, our approach improves the existing st ate-of-the-art methods on the PASCAL VOC 2012 segmentation benchmark, achieving a mIoU score of 66.4% on the test set. Code is available at https://mmcheng.net/ oaa/.
****

Accelerated Gravitational Point Set Alignment With Altered Physical Laws
Vladislav Golyanik, Christian Theobalt, Didier Stricker; Proceedings of the IE EE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2080-2089
This work describes Barnes-Hut Rigid Gravitational Approach (BH-RGA) -- a new ri gid point set registration method relying on principles of particle dynamics. In terpreting the inputs as two interacting particle swarms, we directly minimise t he gravitational potential energy of the system using non-linear least squares. Compared to solutions obtained by solving systems of second-order ordinary diffe rential equations, our approach is more robust and less dependent on the paramet er choice. We accelerate otherwise exhaustive particle interactions with a Barne s-Hut tree and efficiently handle massive point sets in quasilinear time while p reserving the globally multiply-linked character of interactions. Among the adva ntages of BH-RGA is the possibility to define boundary conditions or additional alignment cues through varying point masses. Systematic experiments demonstrate that BH-RGA surpasses performances of baseline methods in terms of the convergen ce basin and accuracy when handling incomplete, noisy and perturbed data. The pr oposed approach also positively compares to the competing method for the alignme nt with prior matches.
****

Domain Adaptation for Semantic Segmentation With Maximum Squares Loss
Minghao Chen, Hongyang Xue, Deng Cai; Proceedings of the IEEE/CVF Internationa l Conference on Computer Vision (ICCV), 2019, pp. 2090-2099
Deep neural networks for semantic segmentation always require a large number of samples with pixel-level labels, which becomes the major difficulty in their rea l-world applications. To reduce the labeling cost, unsupervised domain adaptatio n (UDA) approaches are proposed to transfer knowledge from labeled synthesized d atasets to unlabeled real-world datasets. Recently, some semi-supervised learnin g methods have been applied to UDA and achieved state-of-the-art performance. On e of the most popular approaches in semi-supervised learning is the entropy mini mization method. However, when applying the entropy minimization to UDA for sema ntic segmentation, the gradient of the entropy is biased towards samples that ar e easy to transfer. To balance the gradient of well-classified target samples, w e propose the maximum squares loss. Our maximum squares loss prevents the traini ng process being dominated by easy-to-transfer samples in the target domain. Bes ides, we introduce the image-wise weighting ratio to alleviate the class imbalan ce in the unlabeled target domain. Both synthetic-to-real and cross-city adaptat ion experiments demonstrate the effectiveness of our proposed approach. The code is released at https://github. com/ZJULearning/MaxSquareLoss.
****

Domain Randomization and Pyramid Consistency: Simulation-to-Real Generalization Without Accessing Target Domain Data
Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, Boqing Gong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2100-2110
We propose to harness the potential of simulation for semantic segmentation of r eal-world self-driving scenes in a domain generalization fashion. The segmentati on network is trained without any information about target domains and tested on the unseen target domains. To this end, we propose a new approach of domain ran

domization and pyramid consistency to learn a model with high generalizability. First, we propose to randomize the synthetic images with styles of real images in terms of visual appearances using auxiliary datasets, in order to effectively learn domain-invariant representations. Second, we further enforce pyramid consistency across different "stylized" images and within an image, in order to learn domain-invariant and scale-invariant features, respectively. Extensive experiments are conducted on generalization from GTA and SYNTHIA to Cityscapes, BDDS, and Mapillary; and our method achieves superior results over the state-of-the-art techniques. Remarkably, our generalization results are on par with or even better than those obtained by state-of-the-art simulation-to-real domain adaptation methods, which access the target domain data at training time.
****

Semi-Supervised Skin Detection by Network With Mutual Guidance
Yi He, Jiayuan Shi, Chuan Wang, Haibin Huang, Jiaming Liu, Guanbin Li, Risheng Liu, Jue Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2111-2120
We present a new data-driven method for robust skin detection from a single human portrait image. Unlike previous methods, we incorporate human body as a weak semantic guidance into this task, considering acquiring large-scale of human labeled skin data is commonly expensive and time-consuming. To be specific, we propose a dual-task neural network for joint detection of skin and body via a semi-supervised learning strategy. The dual-task network contains a shared encoder but two decoders for skin and body separately. For each decoder, its output also serves as a guidance for its counterpart, making both decoders mutually guided. Extensive experiments were conducted to demonstrate the effectiveness of our network with mutual guidance, and experimental results show our network outperforms the state-of-the-art in skin detection.
****

ACE: Adapting to Changing Environments for Semantic Segmentation
Zuxuan Wu, Xin Wang, Joseph E. Gonzalez, Tom Goldstein, Larry S. Davis; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2121-2130
Deep neural networks exhibit exceptional accuracy when they are trained and tested on the same data distributions. However, neural classifiers are often extremely brittle when confronted with domain shift---changes in the input distribution that occur over time. We present ACE, a framework for semantic segmentation that dynamically adapts to changing environments over time. By aligning the distribution of labeled training data from the original source domain with the distribution of incoming data in a shifted domain, ACE synthesizes labeled training data for environments as it sees them. This stylized data is then used to update a segmentation model so that it performs well in new environments. To avoid forgetting knowledge from past environments, we introduce a memory that stores feature statistics from previously seen domains. These statistics can be used to replay images in any of the previously observed domains, thus preventing catastrophic forgetting. In addition to standard batch training using stochastic gradient decent (SGD), we also experiment with fast adaptation methods based on adaptive meta-learning. Extensive experiments are conducted on two datasets from SYNTHIA, the results demonstrate the effectiveness of the proposed approach when adapting to a number of tasks.
****

Efficient Segmentation: Learning Downsampling Near Semantic Boundaries
Dmitrii Marin, Zijian He, Peter Vajda, Priyam Chatterjee, Sam Tsai, Fei Yang, Yuri Boykov; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2131-2141
Many automated processes such as auto-piloting rely on a good semantic segmentation as a critical component. To speed up performance, it is common to downsample the input frame. However, this comes at the cost of missed small objects and reduced accuracy at semantic boundaries. To address this problem, we propose a new content-adaptive downsampling technique that learns to favor sampling locations near semantic boundaries of target classes. Cost-performance analysis shows tha

t our method consistently outperforms the uniform sampling improving balance bet
ween accuracy and computational efficiency. Our adaptive sampling gives segmenta
tion with better quality of boundaries and more reliable support for smaller-siz
e objects.
****

Recurrent U-Net for Resource-Constrained Segmentation
Wei Wang, Kaicheng Yu, Joachim Hugonot, Pascal Fua, Mathieu Salzmann; Procee
dings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019,
pp. 2142-2151
State-of-the-art segmentation methods rely on very deep networks that are not al
ways easy to train without very large training datasets and tend to be relativel
y slow to run on standard GPUs. In this paper, we introduce a novel recurrent U-
Net architecture that preserves the compactness of the original U-Net, while sub
stantially increasing its performance to the point where it outperforms the stat
e of the art on several benchmarks. We will demonstrate its effectiveness for se
veral tasks, including hand segmentation, retina vessel segmentation, and road s
egmentation. We also introduce a large-scale dataset for hand segmentation.
****

Detecting the Unexpected via Image Resynthesis
Krzysztof Lis, Krishna Nakka, Pascal Fua, Mathieu Salzmann; Proceedings of th
e IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2152-21
61
Classical semantic segmentation methods, including the recent deep learning ones
, assume that all classes observed at test time have been seen during training.
In this paper, we tackle the more realistic scenario where unexpected objects of
 unknown classes can appear at test time. The main trends in this area either le
verage the notion of prediction uncertainty to flag the regions with low confide
nce as unknown, or rely on autoencoders and highlight poorly-decoded regions. Ha
ving observed that, in both cases, the detected regions typically do not corresp
ond to unexpected objects, in this paper, we introduce a drastically different s
trategy: It relies on the intuition that the network will produce spurious label
s in regions depicting unexpected objects. Therefore, resynthesizing the image f
rom the resulting semantic map will yield significant appearance differences wit
h respect to the input image. In other words, we translate the problem of detect
ing unknown classes to one of identifying poorly-resynthesized image regions. We
 show that this outperforms both uncertainty- and autoencoder-based methods.
****

Self-Supervised Monocular Depth Hints
Jamie Watson, Michael Firman, Gabriel J. Brostow, Daniyar Turmukhambetov; Pro
ceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 201
9, pp. 2162-2171
Monocular depth estimators can be trained with various forms of self-supervision
 from binocular-stereo data to circumvent the need for high-quality laser-scans
or other ground-truth data. The disadvantage, however, is that the photometric r
eprojection losses used with self-supervised learning typically have multiple lo
cal minima. These plausible-looking alternatives to ground-truth can restrict wh
at a regression network learns, causing it to predict depth maps of limited qual
ity. As one prominent example, depth discontinuities around thin structures are
often incorrectly estimated by current state-of-the-art methods. Here, we study
the problem of ambiguous reprojections in depth-prediction from stereo-based sel
f-supervision, and introduce Depth Hints to alleviate their effects. Depth Hints
 are complementary depth-suggestions obtained from simple off-the-shelf stereo a
lgorithms. These hints enhance an existing photometric loss function, and are us
ed to guide a network to learn better weights. They require no additional data,
and are assumed to be right only sometimes. We show that using our Depth Hints g
ives a substantial boost when training several leading self-supervised-from-ster
eo models, not just our own. Further, combined with other good practices, we pro
duce state-of-the-art depth predictions on the KITTI benchmark.
****

3D Scene Reconstruction With Multi-Layer Depth and Epipolar Transformers

Daeyun Shin, Zhile Ren, Erik B. Sudderth, Charless C. Fowlkes; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2172-2182

We tackle the problem of automatically reconstructing a complete 3D model of a scene from a single RGB image. This challenging task requires inferring the shape of both visible and occluded surfaces. Our approach utilizes viewer-centered, multi-layer representation of scene geometry adapted from recent methods for single object shape completion. To improve the accuracy of view-centered representations for complex scenes, we introduce a novel "Epipolar Feature Transformer" that transfers convolutional network features from an input view to other virtual camera viewpoints, and thus better covers the 3D scene geometry. Unlike existing approaches that first detect and localize objects in 3D, and then infer object shape using category-specific models, our approach is fully convolutional, end-to-end differentiable, and avoids the resolution and memory limitations of voxel representations. We demonstrate the advantages of multi-layer depth representations and epipolar feature transformers on the reconstruction of a large database of indoor scenes.

****

How Do Neural Networks See Depth in Single Images?

Tom van Dijk, Guido de Croon; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2183-2191

Deep neural networks have lead to a breakthrough in depth estimation from single images. Recent work shows that the quality of these estimations is rapidly increasing. It is clear that neural networks can see depth in single images. However, to the best of our knowledge, no work currently exists that analyzes what these networks have learned. In this work we take four previously published networks and investigate what depth cues they exploit. We find that all networks ignore the apparent size of known obstacles in favor of their vertical position in the image. The use of the vertical position requires the camera pose to be known; however, we find that these networks only partially recognize changes in camera pitch and roll angles. Small changes in camera pitch are shown to disturb the estimated distance towards obstacles. The use of the vertical image position allows the networks to estimate depth towards arbitrary obstacles - even those not appearing in the training set - but may depend on features that are not universally present.

****

On Boosting Single-Frame 3D Human Pose Estimation via Monocular Videos

Zhi Li, Xuan Wang, Fei Wang, Peilin Jiang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2192-2201

The premise of training an accurate 3D human pose estimation network is the possession of huge amount of richly annotated training data. Nonetheless, manually obtaining rich and accurate annotations is, even not impossible, tedious and slow. In this paper, we propose to exploit monocular videos to complement the training dataset for the single-image 3D human pose estimation tasks. At the beginning, a baseline model is trained with a small set of annotations. By fixing some reliable estimations produced by the resulting model, our method automatically collects the annotations across the entire video as solving the 3D trajectory completion problem. Then, the baseline model is further trained with the collected annotations to learn the new poses. We evaluate our method on the broadly-adopted Human3.6M and MPI-INF-3DHP datasets. As illustrated in experiments, given only a small set of annotations, our method successfully makes the model to learn new poses from unlabelled monocular videos, promoting the accuracies of the baseline model by about 10%. By contrast with previous approaches, our method does not rely on either multi-view imagery or any explicit 2D keypoint annotations.

****

Canonical Surface Mapping via Geometric Cycle Consistency

Nilesh Kulkarni, Abhinav Gupta, Shubham Tulsiani; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2202-2211

We explore the task of Canonical Surface Mapping (CSM). Specifically, given an image, we learn to map pixels on the object to their corresponding locations on a

n abstract 3D model of the category. But how do we learn such a mapping? A supervised approach would require extensive manual labeling which is not scalable beyond a few hand-picked categories. Our key insight is that the CSM task (pixel to 3D), when combined with 3D projection (3D to pixel), completes a cycle. Hence, we can exploit a geometric cycle consistency loss, thereby allowing us to forgo the dense manual supervision. Our approach allows us to train a CSM model for a diverse set of classes, without sparse or dense keypoint annotation, by leveraging only foreground mask labels for training. We show that our predictions also allow us to infer dense correspondence between two images, and compare the performance of our approach against several methods that predict correspondence by leveraging varying amount of supervision.
****

GP2C: Geometric Projection Parameter Consensus for Joint 3D Pose and Focal Length Estimation in the Wild
Alexander Grabner, Peter M. Roth, Vincent Lepetit; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2222-2231
We present a joint 3D pose and focal length estimation approach for object categories in the wild. In contrast to previous methods that predict 3D poses independently of the focal length or assume a constant focal length, we explicitly estimate and integrate the focal length into the 3D pose estimation. For this purpose, we combine deep learning techniques and geometric algorithms in a two-stage approach: First, we estimate an initial focal length and establish 2D-3D correspondences from a single RGB image using a deep network. Second, we recover 3D poses and refine the focal length by minimizing the reprojection error of the predicted correspondences. In this way, we exploit the geometric prior given by the focal length for 3D pose estimation. This results in two advantages: First, we achieve significantly improved 3D translation and 3D pose accuracy compared to existing methods. Second, our approach finds a geometric consensus between the individual projection parameters, which is required for precise 2D-3D alignment. We evaluate our proposed approach on three challenging real-world datasets (Pix3D, Comp, and Stanford) with different object categories and significantly outperform the state-of-the-art by up to 20% absolute in multiple different metrics.
****

Moulding Humans: Non-Parametric 3D Human Shape Estimation From Single Images
Valentin Gabeur, Jean-Sebastien Franco, Xavier Martin, Cordelia Schmid, Gregory Rogez; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2232-2241
In this paper, we tackle the problem of 3D human shape estimation from single RGB images. While the recent progress in convolutional neural networks has allowed impressive results for 3D human pose estimation, estimating the full 3D shape of a person is still an open issue. Model-based approaches can output precise meshes of naked under-cloth human bodies but fail to estimate details and un-modelled elements such as hair or clothing. On the other hand, non-parametric volumetric approaches can potentially estimate complete shapes but, in practice, they are limited by the resolution of the output grid and cannot produce detailed estimates. In this work, we propose a non-parametric approach that employs a double depth map to represent the 3D shape of a person: a visible depth map and a "hidden" depth map are estimated and combined, to reconstruct the human 3D shape as done with a "mould". This representation through 2D depth maps allows a higher resolution output with a much lower dimension than voxel-based volumetric representations. Additionally, our fully derivable depth-based model allows us to efficiently incorporate a discriminator in an adversarial fashion to improve the accuracy and "humanness" of the 3D output. We train and quantitatively validate our approach on SURREAL and on 3D-HUMANS, a new photorealistic dataset made of semi-synthetic in-house videos annotated with 3D ground truth surfaces.
****

3DPeople: Modeling the Geometry of Dressed Humans
Albert Pumarola, Jordi Sanchez-Riera, Gary P. T. Choi, Alberto Sanfeliu, Francesc Moreno-Noguer; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2242-2251

Recent advances in 3D human shape estimation build upon parametric representatio
ns that model very well the shape of the naked body, but are not appropriate to
represent the clothing geometry. In this paper, we present an approach to model
dressed humans and predict their geometry from single images. We contribute in t
hree fundamental aspects of the problem, namely, a new dataset, a novel shape pa
rameterization algorithm and an end-to-end deep generative network for predictin
g shape. First, we present 3DPeople, a large-scale synthetic dataset with 2 Mill
ion photo-realistic images of 80 subjects performing 70 activities and wearing d
iverse outfits. Besides providing textured 3D meshes for clothes and body we ann
otated the dataset with segmentation masks, skeletons, depth, normal maps and op
tical flow. All this together makes 3DPeople suitable for a plethora of tasks. W
e then represent the 3D shapes using 2D geometry images. To build these images w
e propose a novel spherical area-preserving parameterization algorithm based on
the optimal mass transportation method. We show this approach to improve existin
g spherical maps which tend to shrink the elongated parts of the full body model
s such as the arms and legs, making the geometry images incomplete. Finally, we
design a multi-resolution deep generative network that, given an input image of
a dressed human, predicts his/her geometry image (and thus the clothed body shap
e) in an end-to-end manner. We obtain very promising results in jointly capturin
g body pose and clothing shape, both for synthetic validation and on the wild im
ages.

****

Learning to Reconstruct 3D Human Pose and Shape via Model-Fitting in the Loop
Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, Kostas Daniilidis; Pro
ceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 201
9, pp. 2252-2261

Model-based human pose estimation is currently approached through two different
paradigms. Optimization-based methods fit a parametric body model to 2D observat
ions in an iterative manner, leading to accurate image-model alignments, but are
often slow and sensitive to the initialization. In contrast, regression-based m
ethods, that use a deep network to directly estimate the model parameters from p
ixels, tend to provide reasonable, but not pixel accurate, results while requiri
ng huge amounts of supervision. In this work, instead of investigating which app
roach is better, our key insight is that the two paradigms can form a strong col
laboration. A reasonable, directly regressed estimate from the network can initi
alize the iterative optimization making the fitting faster and more accurate. Si
milarly, a pixel accurate fit from iterative optimization can act as strong supe
rvision for the network. This is the core of our proposed approach SPIN (SMPL oP
timization IN the loop). The deep network initializes an iterative optimization
routine that fits the body model to 2D joints within the training loop, and the
fitted estimate is subsequently used to supervise the network. Our approach is s
elf-improving by nature, since better network estimates can lead the optimizatio
n to better solutions, while more accurate optimization fits provide better supe
rvision for the network. We demonstrate the effectiveness of our approach in dif
ferent settings, where 3D ground truth is scarce, or not available, and we consi
stently outperform the state-of-the-art model-based pose estimation approaches b
y significant margins. The project website with videos, results, and code can be
found at https://seas.upenn.edu/ nkolot/projects/spin.

****

Optimizing Network Structure for 3D Human Pose Estimation
Hai Ci, Chunyu Wang, Xiaoxuan Ma, Yizhou Wang; Proceedings of the IEEE/CVF In
ternational Conference on Computer Vision (ICCV), 2019, pp. 2262-2271

A human pose is naturally represented as a graph where the joints are the nodes
and the bones are the edges. So it is natural to apply Graph Convolutional Netwo
rk (GCN) to estimate 3D poses from 2D poses. In this work, we propose a generic
formulation where both GCN and Fully Connected Network (FCN) are its special cas
es. From this formulation, we discover that GCN has limited representation power
when used for estimating 3D poses. We overcome the limitation by introducing Lo
cally Connected Network (LCN) which is naturally implemented by this generic for
mulation. It notably improves the representation capability over GCN. In additio

n, since every joint is only connected to a few joints in its neighborhood, it h as strong generalization power. The experiments on public datasets show it: (1) outperforms the state-of-the-arts; (2) is less data hungry than alternative mode ls; (3) generalizes well to unseen actions and datasets.
****

Exploiting Spatial-Temporal Relationships for 3D Pose Estimation via Graph Convo lutional Networks
Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, Na dia Magnenat Thalmann; Proceedings of the IEEE/CVF International Conference on C omputer Vision (ICCV), 2019, pp. 2272-2281
Despite great progress in 3D pose estimation from single-view images or videos, it remains a challenging task due to the substantial depth ambiguity and severe self-occlusions. Motivated by the effectiveness of incorporating spatial depende ncies and temporal consistencies to alleviate these issues, we propose a novel g raph-based method to tackle the problem of 3D human body and 3D hand pose estima tion from a short sequence of 2D joint detections. Particularly, domain knowledg e about the human hand (body) configurations is explicitly incorporated into the graph convolutional operations to meet the specific demand of the 3D pose estim ation. Furthermore, we introduce a local-to-global network architecture, which i s capable of learning multi-scale features for the graph-based representations. We evaluate the proposed method on challenging benchmark datasets for both 3D ha nd pose estimation and 3D body pose estimation. Experimental results show that o ur method achieves state-of-the-art performance on both tasks.
****

Resolving 3D Human Pose Ambiguities With 3D Scene Constraints
Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, Michael J. Black; Proce edings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2282-2292
To understand and analyze human behavior, we need to capture humans moving in, a nd interacting with, the world. Most existing methods perform 3D human pose esti mation without explicitly considering the scene. We observe however that the wor ld constrains the body and vice-versa. To motivate this, we show that current 3D human pose estimation methods produce results that are not consistent with the 3D scene. Our key contribution is to exploit static 3D scene structure to better estimate human pose from monocular images. The method enforces Proximal Relatio nships with Object eXclusion and is called PROX. To test this, we collect a new dataset composed of 12 different 3D scenes and RGB sequences of 20 subjects movi ng in and interacting with the scenes. We represent human pose using the 3D huma n body model SMPL-X and extend SMPLify-X to estimate body pose using scene const raints. We make use of the 3D scene information by formulating two main constrai nts. The inter-penetration constraint penalizes intersection between the body mo del and the surrounding 3D scene. The contact constraint encourages specific par ts of the body to be in contact with scene surfaces if they are close enough in distance and orientation. For quantitative evaluation we capture a separate data set with 180 RGB frames in which the ground-truth body pose is estimated using a motion capture system. We show quantitatively that introducing scene constraint s significantly reduces 3D joint error and vertex error. Our code and data are a vailable for research at https://prox.is.tue.mpg.de.
****

Tex2Shape: Detailed Full Human Body Geometry From a Single Image
Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, Marcus Magnor; Proceed ings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, p p. 2293-2303
We present a simple yet effective method to infer detailed full human body shape from only a single photograph. Our model can infer full-body shape including fa ce, hair, and clothing including wrinkles at interactive frame-rates. Results fe ature details even on parts that are occluded in the input image. Our main idea is to turn shape regression into an aligned image-to-image translation problem. The input to our method is a partial texture map of the visible region obtained from off-the-shelf methods. From a partial texture, we estimate detailed normal

and vector displacement maps, which can be applied to a low-resolution smooth body model to add detail and clothing. Despite being trained purely with synthetic data, our model generalizes well to real-world photographs. Numerous results demonstrate the versatility and robustness of our method.
****

PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization
Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, Hao Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2304-2314
We introduce Pixel-aligned Implicit Function (PIFu), an implicit representation that locally aligns pixels of 2D images with the global context of their corresponding 3D object. Using PIFu, we propose an end-to-end deep learning method for digitizing highly detailed clothed humans that can infer both 3D surface and texture from a single image, and optionally, multiple input images. Highly intricate shapes, such as hairstyles, clothing, as well as their variations and deformations can be digitized in a unified way. Compared to existing representations used for 3D deep learning, PIFu produces high-resolution surfaces including largely unseen regions such as the back of a person. In particular, it is memory efficient unlike the voxel representation, can handle arbitrary topology, and the resulting surface is spatially aligned with the input image. Furthermore, while previous techniques are designed to process either a single image or multiple views, PIFu extends naturally to arbitrary number of views. We demonstrate high-resolution and robust reconstructions on real world images from the DeepFashion dataset, which contains a variety of challenging clothing types. Our method achieves state-of-the-art performance on a public benchmark and outperforms the prior work for clothed human digitization from a single image.
****

DF2Net: A Dense-Fine-Finer Network for Detailed 3D Face Reconstruction
Xiaoxing Zeng, Xiaojiang Peng, Yu Qiao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2315-2324
Reconstructing the detailed geometric structure from a single face image is a challenging problem due to its ill-posed nature and the fine 3D structures to be recovered. This paper proposes a deep Dense-Fine-Finer Network (DF2Net) to address this challenging problem. DF2Net decomposes the reconstruction process into three stages, each of which is processed by an elaborately-designed network, namely D-Net, F-Net, and Fr-Net. D-Net exploits a U-net architecture to map the input image to a dense depth image. F-Net refines the output of D-Net by integrating features from depth and RGB domains, whose output is further enhanced by Fr-Net with a novel multi-resolution hypercolumn architecture. In addition, we introduce three types of data to train these networks, including 3D model synthetic data, 2D image reconstructed data, and fine facial images. We elaborately exploit different datasets (or combination) together with well-designed losses to train different networks. Qualitative evaluation indicates that our DF2Net can effectively reconstruct subtle facial details such as small crow's feet and wrinkles. Our DF2Net achieves performance superior or comparable to state-of-the-art algorithms in qualitative and quantitative analyses on real-world images and the BU-3DFE dataset. Code and the collected 70K image-depth data will be publicly available.
****

Monocular 3D Human Pose Estimation by Generation and Ordinal Ranking
Saurabh Sharma, Pavan Teja Varigonda, Prashast Bindal, Abhishek Sharma, Arjun Jain; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2325-2334
Monocular 3D human-pose estimation from static images is a challenging problem, due to the curse of dimensionality and the ill-posed nature of lifting 2D-to-3D. In this paper, we propose a Deep Conditional Variational Autoencoder based model that synthesizes diverse anatomically plausible 3D-pose samples conditioned on the estimated 2D-pose. We show that CVAE-based 3D-pose sample set is consistent with the 2D-pose and helps tackling the inherent ambiguity in 2D-to-3D lifting.

We propose two strategies for obtaining the final 3D pose- (a) depth-ordering/ordinal relations to score and weight-average the candidate 3D-poses, referred to as OrdinalScore, and (b) with supervision from an Oracle. We report close to state-of-the-art results on two benchmark datasets using OrdinalScore, and state-of-the-art results using the Oracle. We also show that our pipeline yields competitive results without paired image-to-3D annotations. The training and evaluation code is available at https://github.com/ssfootball04/generative_pose.
****

Aligning Latent Spaces for 3D Hand Pose Estimation
Linlin Yang, Shile Li, Dongheui Lee, Angela Yao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2335-2343
Hand pose estimation from monocular RGB inputs is a highly challenging task. Many previous works for monocular settings only used RGB information for training despite the availability of corresponding data in other modalities such as depth maps. In this work, we propose to learn a joint latent representation that leverages other modalities as weak labels to boost the RGB-based hand pose estimator. By design, our architecture is highly flexible in embedding various diverse modalities such as heat maps, depth maps and point clouds. In particular, we find that encoding and decoding the point cloud of the hand surface can improve the quality of the joint latent representation. Experiments show that with the aid of other modalities during training, our proposed method boosts the accuracy of RGB-based hand pose estimation systems and significantly outperforms state-of-the-art on two public benchmarks.
****

HEMlets Pose: Learning Part-Centric Heatmap Triplets for Accurate 3D Human Pose Estimation
Kun Zhou, Xiaoguang Han, Nianjuan Jiang, Kui Jia, Jiangbo Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2344-2353
Estimating 3D human pose from a single image is a challenging task. This work attempts to address the uncertainty of lifting the detected 2D joints to the 3D space by introducing an intermediate state - Part-Centric Heatmap Triplets (HEMlets), which shortens the gap between the 2D observation and the 3D interpretation. The HEMlets utilize three joint-heatmaps to represent the relative depth information of the end-joints for each skeletal body part. In our approach, a Convolutional Network(ConvNet) is first trained to predict HEMlests from the input image, followed by a volumetric joint-heatmap regression. We leverage on the integral operation to extract the joint locations from the volumetric heatmaps, guaranteeing end-to-end learning. Despite the simplicity of the network design, the quantitative comparisons show a significant performance improvement over the best-of-grade method (by 20% on Human3.6M). The proposed method naturally supports training with "in-the-wild" images, where only weakly-annotated relative depth information of skeletal joints is available. This further improves the generalization ability of our model, as validated by qualitative comparisons on outdoor images.
****

End-to-End Hand Mesh Recovery From a Monocular RGB Image
Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, Wen Zheng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2354-2364
In this paper, we present a HAnd Mesh Recovery (HAMR) framework to tackle the problem of reconstructing the full 3D mesh of a human hand from a single RGB image. In contrast to existing research on 2D or 3D hand pose estimation from RGB or/and depth image data, HAMR can provide a more expressive and useful mesh representation for monocular hand image understanding. In particular, the mesh representation is achieved by parameterizing a generic 3D hand model with shape and relative 3D joint angles. By utilizing this mesh representation, we can easily compute the 3D joint locations via linear interpolations between the vertexes of the mesh, while obtain the 2D joint locations with a projection of the 3D joints. To this end, a differentiable re-projection loss can be defined in terms of the derived representations and the ground-truth labels, thus making our framework end

-to-end trainable. Qualitative experiments show that our framework is capable of recovering appealing 3D hand mesh even in the presence of severe occlusions. Quantitatively, our approach also outperforms the state-of-the-art methods for both 2D and 3D hand pose estimation from a monocular RGB image on several benchmark datasets.

****

Robust Multi-Modality Multi-Object Tracking

Wenwei Zhang, Hui Zhou, Shuyang Sun, Zhe Wang, Jianping Shi, Chen Change Loy; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2365-2374

Multi-sensor perception is crucial to ensure the reliability and accuracy in autonomous driving system, while multi-object tracking (MOT) improves that by tracing sequential movement of dynamic objects. Most current approaches for multi-sensor multi-object tracking are either lack of reliability by tightly relying on a single input source (e.g., center camera), or not accurate enough by fusing the results from multiple sensors in post processing without fully exploiting the inherent information. In this study, we design a generic sensor-agnostic multi-modality MOT framework (mmMOT), where each modality (i.e., sensors) is capable of performing its role independently to preserve reliability, and could further improving its accuracy through a novel multi-modality fusion module. Our mmMOT can be trained in an end-to-end manner, enables joint optimization for the base feature extractor of each modality and an adjacency estimator for cross modality. Our mmMOT also makes the first attempt to encode deep representation of point cloud in data association process in MOT. We conduct extensive experiments to evaluate the effectiveness of the proposed framework on the challenging KITTI benchmark and report state-of-the-art performance. Code and models are available at https://github.com/ZwwWayne/mmMOT.

****

The Trajectron: Probabilistic Multi-Agent Trajectory Modeling With Dynamic Spatiotemporal Graphs

Boris Ivanovic, Marco Pavone; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2375-2384

Developing safe human-robot interaction systems is a necessary step towards the widespread integration of autonomous agents in society. A key component of such systems is the ability to reason about the many potential futures (e.g. trajectories) of other agents in the scene. Towards this end, we present the Trajectron, a graph-structured model that predicts many potential future trajectories of multiple agents simultaneously in both highly dynamic and multimodal scenarios (i.e. where the number of agents in the scene is time-varying and there are many possible highly-distinct futures for each agent). It combines tools from recurrent sequence modeling and variational deep generative modeling to produce a distribution of future trajectories for each agent in a scene. We demonstrate the performance of our model on several datasets, obtaining state-of-the-art results on standard trajectory prediction metrics as well as introducing a new metric for comparing models that output distributions.

****

'Skimming-Perusal' Tracking: A Framework for Real-Time and Robust Long-Term Tracking

Bin Yan, Haojie Zhao, Dong Wang, Huchuan Lu, Xiaoyun Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2385-2393

Compared with traditional short-term tracking, long-term tracking poses more challenges and is much closer to realistic applications. However, few works have been done and their performance have also been limited. In this work, we present a novel robust and real-time long-term tracking framework based on the proposed skimming and perusal modules. The perusal module consists of an effective bounding box regressor to generate a series of candidate proposals and a robust target verifier to infer the optimal candidate with its confidence score. Based on this score, our tracker determines whether the tracked object being present or absent, and then chooses the tracking strategies of local search or global search res

pectively in the next frame. To speed up the image-wide global search, a novel s kimming module is designed to efficiently choose the most possible regions from a large number of sliding windows. Numerous experimental results on the VOT-2018 long-term and OxUvA long-term benchmarks demonstrate that the proposed method a chieves the best performance and runs in real-time. The source codes are availab le at https://github.com/iiau-tracker/SPLT.
****

TASED-Net: Temporally-Aggregating Spatial Encoder-Decoder Network for Video Sali ency Detection
Kyle Min, Jason J. Corso; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2394-2403
TASED-Net is a 3D fully-convolutional network architecture for video saliency de tection. It consists of two building blocks: first, the encoder network extracts low-resolution spatiotemporal features from an input clip of several consecutiv e frames, and then the following prediction network decodes the encoded features spatially while aggregating all the temporal information. As a result, a single prediction map is produced from an input clip of multiple frames. Frame-wise sa liency maps can be predicted by applying TASED-Net in a sliding-window fashion t o a video. The proposed approach assumes that the saliency map of any frame can be predicted by considering a limited number of past frames. The results of our extensive experiments on video saliency detection validate this assumption and d emonstrate that our fully-convolutional model with temporal aggregation method i s effective. TASED-Net significantly outperforms previous state-of-the-art appro aches on all three major large-scale datasets of video saliency detection: DHF1K , Hollywood2, and UCFSports. After analyzing the results qualitatively, we obser ve that our model is especially better at attending to salient moving objects.
****

Attacking Optical Flow
Anurag Ranjan, Joel Janai, Andreas Geiger, Michael J. Black; Proceedings of t he IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2404-2 413
Deep neural nets achieve state-of-the-art performance on the problem of optical flow estimation. Since optical flow is used in several safety-critical applicati ons like self-driving cars, it is important to gain insights into the robustness of those techniques. Recently, it has been shown that adversarial attacks easil y fool deep neural networks to misclassify objects. The robustness of optical fl ow networks to adversarial attacks, however, has not been studied so far. In thi s paper, we extend adversarial patch attacks to optical flow networks and show t hat such attacks can compromise their performance. We show that corrupting a sma ll patch of less than 1% of the image size can significantly affect optical flow estimates. Our attacks lead to noisy flow estimates that extend significantly b eyond the region of the attack, in many cases even completely erasing the motion of objects in the scene. While networks using an encoder-decoder architecture a re very sensitive to these attacks, we found that networks using a spatial pyram id architecture are less affected. We analyse the success and failure of attacki ng both architectures by visualizing their feature maps and comparing them to cl assical optical flow techniques which are robust to these attacks. We also demon strate that such attacks are practical by placing a printed pattern into real sc enes.
****

Pro-Cam SSfM: Projector-Camera System for Structure and Spectral Reflectance Fro m Motion
Chunyu Li, Yusuke Monno, Hironori Hidaka, Masatoshi Okutomi; Proceedings of t he IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2414-2 423
In this paper, we propose a novel projector-camera system for practical and low- cost acquisition of a dense object 3D model with the spectral reflectance proper ty. In our system, we use a standard RGB camera and leverage an off-the-shelf pr ojector as active illumination for both the 3D reconstruction and the spectral r eflectance estimation. We first reconstruct the 3D points while estimating the p

oses of the camera and the projector, which are alternately moved around the obj ect, by combining multi-view structured light and structure-from-motion (SfM) te chniques. We then exploit the projector for multispectral imaging and estimate t he spectral reflectance of each 3D point based on a novel spectral reflectance e stimation model considering the geometric relationship between the reconstructed 3D points and the estimated projector positions. Experimental results on severa l real objects demonstrate that our system can precisely acquire a dense 3D mode l with the full spectral reflectance property using off-the-shelf devices.
****

Mop Moire Patterns Using MopNet
Bin He, Ce Wang, Boxin Shi, Ling-Yu Duan; Proceedings of the IEEE/CVF Interna tional Conference on Computer Vision (ICCV), 2019, pp. 2424-2432
Moire pattern is a common image quality degradation caused by frequency aliasing between monitors and cameras when taking screen-shot photos. The complex freque ncy distribution, imbalanced magnitude in colour channels, and diverse appearanc e attributes of moire pattern make its removal a challenging problem. In this pa per, we propose a Moire pattern Removal Neural Network (MopNet) to solve this pr oblem. All core components of MopNet are specially designed for unique propertie s of moire patterns, including the multi-scale feature aggregation addressing co mplex frequency, the channel-wise target edge predictor to exploit imbalanced ma gnitude among colour channels, and the attribute-aware classifier to characteriz e the diverse appearance for better modelling Moire patterns. Quantitative and q ualitative experimental comparison validate the state-of-the-art performance of MopNet.
****

Kernel Modeling Super-Resolution on Real Low-Resolution Images
Ruofan Zhou, Sabine Susstrunk; Proceedings of the IEEE/CVF International Confer ence on Computer Vision (ICCV), 2019, pp. 2433-2443
Deep convolutional neural networks (CNNs), trained on corresponding pairs of hig h- and low-resolution images, achieve state-of-the-art performance in single-ima ge super-resolution and surpass previous signal-processing based approaches. How ever, their performance is limited when applied to real photographs. The reason lies in their training data: low-resolution (LR) images are obtained by bicubic interpolation of the corresponding high-resolution (HR) images. The applied conv olution kernel significantly differs from real-world camera-blur. Consequently, while current CNNs well super-resolve bicubic-downsampled LR images, they often fail on camera-captured LR images. To improve generalization and robustness of d eep super-resolution CNNs on real photographs, we present a kernel modeling supe r-resolution network (KMSR) that incorporates blur-kernel modeling in the traini ng. Our proposed KMSR consists of two stages: we first build a pool of realistic blur-kernels with a generative adversarial network (GAN) and then we train a su per-resolution network with HR and corresponding LR images constructed with the generated kernels. Our extensive experimental validations demonstrate the effect iveness of our single-image super-resolution approach on photographs with unknow n blur-kernels.
****

Learning to Jointly Generate and Separate Reflections
Daiqian Ma, Renjie Wan, Boxin Shi, Alex C. Kot, Ling-Yu Duan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2444 -2452
Existing learning-based single image reflection removal methods using paired tra ining data have fundamental limitations about the generalization capability on r eal-world reflections due to the limited variations in training pairs. In this w ork, we propose to jointly generate and separate reflections within a weakly-sup ervised learning framework, aiming to model the reflection image formation more comprehensively with abundant unpaired supervision. By imposing the adversarial losses and combinable mapping mechanism in a multi-task structure, the proposed framework elegantly integrates the two separate stages of reflection generation and separation into a unified model. The gradient constraint is incorporated int o the concurrent training process of the multi-task learning as well. In particu

lar, we built up an unpaired reflection dataset with 4,027 images, which is useful for facilitating the weakly-supervised learning of reflection removal model. Extensive experiments on a public benchmark dataset show that our framework performs favorably against state-of-the-art methods and consistently produces visually appealing results.
****

Deep Multi-Model Fusion for Single-Image Dehazing
Zijun Deng, Lei Zhu, Xiaowei Hu, Chi-Wing Fu, Xuemiao Xu, Qing Zhang, Jing Qin, Pheng-Ann Heng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2453-2462
This paper presents a deep multi-model fusion network to attentively integrate multiple models to separate layers and boost the performance in single-image dehazing. To do so, we first formulate the attentional feature integration module to maximize the integration of the convolutional neural network (CNN) features at different CNN layers and generate the attentional multi-level integrated features (AMLIF). Then, from the AMLIF, we further predict a haze-free result for an atmospheric scattering model, as well as for four haze-layer separation models, and then fuse the results together to produce the final haze-free image. To evaluate the effectiveness of our method, we compare our network with several state-of-the-art methods on two widely-used dehazing benchmark datasets, as well as on two sets of real-world hazy images. Experimental results demonstrate clear quantitative and qualitative improvements of our method over the state-of-the-arts.
****

Deep Learning for Seeing Through Window With Raindrops
Yuhui Quan, Shijie Deng, Yixin Chen, Hui Ji; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2463-2471
When taking pictures through glass window in rainy day, the images are comprised and corrupted by the raindrops adhered to glass surfaces. It is a challenging problem to remove the effect of raindrops from an image. The key task is how to accurately and robustly identify the raindrop regions in an image. This paper develops a convolutional neural network (CNN) for removing the effect of raindrops from an image. In the proposed CNN, we introduce a double attention mechanism that concurrently guides the CNN using shape-driven attention and channel re-calibration. The shape-driven attention exploits physical shape priors of raindrops, i.e. convexness and contour closedness, to accurately locate raindrops, and the channel re-calibration improves the robustness when processing raindrops with varying appearances. The experimental results show that the proposed CNN outperforms the state-of-the-art approaches in terms of both quantitative metrics and visual quality.
****

Mask-ShadowGAN: Learning to Remove Shadows From Unpaired Data
Xiaowei Hu, Yitong Jiang, Chi-Wing Fu, Pheng-Ann Heng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2472-2481
This paper presents a new method for shadow removal using unpaired data, enabling us to avoid tedious annotations and obtain more diverse training samples. However, directly employing adversarial learning and cycle-consistency constraints is insufficient to learn the underlying relationship between the shadow and shadow-free domains, since the mapping between shadow and shadow-free images is not simply one-to-one. To address the problem, we formulate Mask-ShadowGAN, a new deep framework that automatically learns to produce a shadow mask from the input shadow image and then takes the mask to guide the shadow generation via re-formulated cycle-consistency constraints. Particularly, the framework simultaneously learns to produce shadow masks and learns to remove shadows, to maximize the overall performance. Also, we prepared an unpaired dataset for shadow removal and demonstrated the effectiveness of Mask-ShadowGAN on various experiments, even it was trained on unpaired data.
****

Spatio-Temporal Filter Adaptive Network for Video Deblurring
Shangchen Zhou, Jiawei Zhang, Jinshan Pan, Haozhe Xie, Wangmeng Zuo, Jimmy Ren; Proceedings of the IEEE/CVF International Conference on Computer Vision (IC

CV), 2019, pp. 2482-2491
Video deblurring is a challenging task due to the spatially variant blur caused by camera shake, object motions, and depth variations, etc. Existing methods usually estimate optical flow in the blurry video to align consecutive frames or approximate blur kernels. However, they tend to generate artifacts or cannot effectively remove blur when the estimated optical flow is not accurate. To overcome the limitation of separate optical flow estimation, we propose a Spatio-Temporal Filter Adaptive Network (STFAN) for the alignment and deblurring in a unified framework. The proposed STFAN takes both blurry and restored images of the previous frame as well as blurry image of the current frame as input, and dynamically generates the spatially adaptive filters for the alignment and deblurring. We then propose the new Filter Adaptive Convolutional (FAC) layer to align the deblurred features of the previous frame with the current frame and remove the spatially variant blur from the features of the current frame. Finally, we develop a reconstruction network which takes the fusion of two transformed features to restore the clear frames. Both quantitative and qualitative evaluation results on the benchmark datasets and real-world videos demonstrate that the proposed algorithm performs favorably against state-of-the-art methods in terms of accuracy, speed as well as model size.
****

Learning Deep Priors for Image Dehazing
Yang Liu, Jinshan Pan, Jimmy Ren, Zhixun Su; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2492-2500
Image dehazing is a well-known ill-posed problem, which usually requires some image priors to make the problem well-posed. We propose an effective iteration algorithm with deep CNNs to learn haze-relevant priors for image dehazing. We formulate the image dehazing problem as the minimization of a variational model with favorable data fidelity terms and prior terms to regularize the model. We solve the variational model based on the classical gradient descent method with built-in deep CNNs so that iteration-wise image priors for the atmospheric light, transmission map and clear image can be well estimated. Our method combines the properties of both the physical formation of image dehazing as well as deep learning approaches. We show that it is able to generate clear images as well as accurate atmospheric light and transmission maps. Extensive experimental results demonstrate that the proposed algorithm performs favorably against state-of-the-art methods in both benchmark datasets and real-world images.
****

JPEG Artifacts Reduction via Deep Convolutional Sparse Coding
Xueyang Fu, Zheng-Jun Zha, Feng Wu, Xinghao Ding, John Paisley; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2501-2510
To effectively reduce JPEG compression artifacts, we propose a deep convolutional sparse coding (DCSC) network architecture. We design our DCSC in the framework of classic learned iterative shrinkage-threshold algorithm. To focus on recognizing and separating artifacts only, we sparsely code the feature maps instead of the raw image. The final de-blocked image is directly reconstructed from the coded features. We use dilated convolution to extract multi-scale image features, which allows our single model to simultaneously handle multiple JPEG compression levels. Since our method integrates model-based convolutional sparse coding with a learning-based deep neural network, the entire network structure is compact and more explainable. The resulting lightweight model generates comparable or better de-blocking results when compared with state-of-the-art methods.
****

Self-Guided Network for Fast Image Denoising
Shuhang Gu, Yawei Li, Luc Van Gool, Radu Timofte; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2511-2520
During the past years, tremendous advances in image restoration tasks have been achieved using highly complex neural networks. Despite their good restoration performance, the heavy computational burden hinders the deployment of these networks on constrained devices, e.g. smart phones and consumer electronic products. T

o tackle this problem, we propose a self-guided network (SGN), which adopts a to p-down self-guidance architecture to better exploit image multi-scale informatio n. SGN directly generates multi-resolution inputs with the shuffling operation. Large-scale contextual information extracted at low resolution is gradually prop agated into the higher resolution sub-networks to guide the feature extraction p rocesses at these scales. Such a self-guidance strategy enables SGN to efficient ly incorporate multi-scale information and extract good local features to recove r noisy images. We validate the effectiveness of SGN through extensive experimen ts. The experimental results demonstrate that SGN greatly improves the memory an d runtime efficiency over state-of-the-art efficient methods, without trading of f PSNR accuracy.

****

Non-Local Intrinsic Decomposition With Near-Infrared Priors

Ziang Cheng, Yinqiang Zheng, Shaodi You, Imari Sato; Proceedings of the IEEE/ CVF International Conference on Computer Vision (ICCV), 2019, pp. 2521-2530

Intrinsic image decomposition is a highly under-constrained problem that has bee n extensively studied by computer vision researchers. Previous methods impose ad ditional constraints by exploiting either empirical or data-driven priors. In th is paper, we revisit intrinsic image decomposition with the aid of near-infrared (NIR) imagery. We show that NIR band is considerably less sensitive to textures and can be exploited to reduce ambiguity caused by reflectance variation, promo ting a simple yet powerful prior for shading smoothness. With this observation, we formulate intrinsic decomposition as an energy minimisation problem. Unlike e xisting methods, our energy formulation decouples reflectance and shading estima tion, into a convex local shading component based on NIR-RGB image pair, and a r eflectance component that encourages reflectance homogeneity both locally and gl obally. We further show the minimisation process can be approached by a series o f multi-dimensional kernel convolutions, each within linear time complexity. To validate the proposed algorithm, a NIR-RGB dataset is captured over real-world o bjects, where our NIR-assisted approach demonstrates clear superiority over RGB methods.

****

VideoMem: Constructing, Analyzing, Predicting Short-Term and Long-Term Video Mem orability

Romain Cohendet, Claire-Helene Demarty, Ngoc Q. K. Duong, Martin Engilberge; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2531-2540

Humans share a strong tendency to memorize/forget some of the visual information they encounter. This paper focuses on understanding the intrinsic memorability of visual content. To address this challenge, we introduce a large scale dataset (VideoMem) composed of 10,000 videos with memorability scores. In contrast to p revious work on image memorability -- where memorability was measured a few minu tes after memorization -- memory performance is measured twice: a few minutes an d again 24-72 hours after memorization. Hence, the dataset comes with short-term and long-term memorability annotations. After an in-depth analysis of the datas et, we investigate various deep neural network-based models for the prediction o f video memorability. Our best model using a ranking loss achieves a Spearman's rank correlation of 0.494 (respectively 0.256) for short-term (resp. long-term) memorability prediction, while our model with attention mechanism provides insig hts of what makes a content memorable. The VideoMem dataset with pre-extracted f eatures is publicly available.

****

Rescan: Inductive Instance Segmentation for Indoor RGBD Scans

Maciej Halber, Yifei Shi, Kai Xu, Thomas Funkhouser; Proceedings of the IEEE/ CVF International Conference on Computer Vision (ICCV), 2019, pp. 2541-2550

In depth-sensing applications ranging from home robotics to AR/VR, it will be co mmon to acquire 3D scans of interior spaces repeatedly at sparse time intervals (e.g., as part of regular daily use). We propose an algorithm that analyzes thes e "rescans" to infer a temporal model of a scene with semantic instance informat ion. Our algorithm operates inductively by using the temporal model resulting fr

om past observations to infer an instance segmentation of a new scan, which is t
hen used to update the temporal model. The model contains object instance associ
ations across time and thus can be used to track individual objects, even though
 there are only sparse observations. During experiments with a new benchmark for
 the new task, our algorithm outperforms alternate approaches based on state-of-
the-art networks for semantic instance segmentation.
****

End-to-End CAD Model Retrieval and 9DoF Alignment in 3D Scans
Armen Avetisyan, Angela Dai, Matthias Niessner; Proceedings of the IEEE/CVF In
ternational Conference on Computer Vision (ICCV), 2019, pp. 2551-2560
We present a novel, end-to-end approach to align CAD models to an 3D scan of a s
cene, enabling transformation of a noisy, incomplete 3D scan to a compact, CAD r
econstruction with clean, complete object geometry. Our main contribution lies i
n formulating a differentiable Procrustes alignment that is paired with a symmet
ry-aware dense object correspondence prediction. To simultaneously align CAD mod
els to all the objects of a scanned scene, our approach detects object locations
, then predicts symmetry-aware dense object correspondences between scan and CAD
 geometry in a unified object space, as well as a nearest neighbor CAD model, bo
th of which are then used to inform a differentiable Procrustes alignment. Our a
pproach operates in a fully-convolutional fashion, enabling alignment of CAD mod
els to the objects of a scan in a single forward pass. This enables our method t
o outperform state-of-the-art approaches by 19.04% for CAD model alignment to sc
ans, with approximately 250x faster runtime than previous data-driven approaches
.
****

Making History Matter: History-Advantage Sequence Training for Visual Dialog
Tianhao Yang, Zheng-Jun Zha, Hanwang Zhang; Proceedings of the IEEE/CVF Intern
ational Conference on Computer Vision (ICCV), 2019, pp. 2561-2569
We study the multi-round response generation in visual dialog, where a response
 is generated according to a visually grounded conversational history. Given a tr
iplet: an image, Q&A history, and current question, all the prevailing methods f
ollow a codec (i.e., encoder-decoder) fashion in a supervised learning paradigm:
 a multimodal encoder encodes the triplet into a feature vector, which is then f
ed into the decoder for the current answer generation, supervised by the ground-
truth. However, this conventional supervised learning does NOT take into account
 the impact of imperfect history, violating the conversational nature of visual
dialog and thus making the codec more inclined to learn history bias but not con
textual reasoning. To this end, inspired by the actor-critic policy gradient in
reinforcement learning, we propose a novel training paradigm called History Adva
ntage Sequence Training (HAST). Specifically, we intentionally impose wrong answ
ers in the history, obtaining an adverse critic, and see how the historic error
impacts the codec's future behavior by History Advantage -- a quantity obtained
by subtracting the adverse critic from the gold reward of ground-truth history.
Moreover, to make the codec more sensitive to the history, we propose a novel at
tention network called History-Aware Co-Attention Network (HACAN) which can be e
ffectively trained by using HAST. Experimental results on three benchmarks: VisD
ial v0.9&v1.0 and GuessWhat?!, show that the proposed HAST strategy consistently
 outperforms the state-of-the-art supervised counterparts.
****

Stochastic Attraction-Repulsion Embedding for Large Scale Image Localization
Liu Liu, Hongdong Li, Yuchao Dai; Proceedings of the IEEE/CVF International Co
nference on Computer Vision (ICCV), 2019, pp. 2570-2579
This paper tackles the problem of large-scale image-based localization (IBL) whe
re the spatial location of a query image is determined by finding out the most s
imilar reference images in a large database. For solving this problem, a critica
l task is to learn discriminative image representation that captures informative
 information relevant for localization. We propose a novel representation learni
ng method having higher location-discriminating power. It provides the following
 contributions: 1) we represent a place (location) as a set of exemplar images d
epicting the same landmarks and aim to maximize similarities among intra-place i

mages while minimizing similarities among inter-place images; 2) we model a simi
larity measure as a probability distribution on L_2-metric distances between int
ra-place and inter-place image representations; 3) we propose a new Stochastic A
ttraction and Repulsion Embedding (SARE) loss function minimizing the KL diverge
nce between the learned and the actual probability distributions; 4) we give the
oretical comparisons between SARE, triplet ranking and contrastive losses. It pr
ovides insights into why SARE is better by analyzing gradients. Our SARE loss is
 easy to implement and pluggable to any CNN. Experiments show that our proposed
method improves the localization performance on standard benchmarks by a large m
argin. Demonstrating the broad applicability of our method, we obtained the thir
d place out of 209 teams in the 2018 Google Landmark Retrieval Challenge. Our co
de and model are available at https://github.com/Liumouliu/deepIBL.
****

Scene Graph Prediction With Limited Labels

Vincent S. Chen, Paroma Varma, Ranjay Krishna, Michael Bernstein, Christophe
r Re, Li Fei-Fei; Proceedings of the IEEE/CVF International Conference on Compu
ter Vision (ICCV), 2019, pp. 2580-2590

Visual knowledge bases such as Visual Genome power numerous applications in comp
uter vision, including visual question answering and captioning, but suffer from
 sparse, incomplete relationships. All scene graph models to date are limited to
 training on a small set of visual relationships that have thousands of training
 labels each. Hiring human annotators is expensive, and using textual knowledge
base completion methods are incompatible with visual data. In this paper, we int
roduce a semi-supervised method that assigns probabilistic relationship labels t
o a large number of unlabeled images using few labeled examples. We analyze visu
al relationships to suggest two types of image-agnostic features that are used t
o generate noisy heuristics, whose outputs are aggregated using a factor graph-b
ased generative model. With as few as 10 labeled examples per relationship, the
generative model creates enough training data to train any existing state-of-the
-art scene graph model. We demonstrate that our method outperforms all baseline
approaches on scene graph prediction by5.16 recall@100 for PREDCLS. In our limit
ed label setting, we define a complexity metric for relationships that serves as
 an indicator (R^2 = 0.778) for conditions under which our method succeeds over
transfer learning, the de-facto approach for training with limited labels.
****

Taking a HINT: Leveraging Explanations to Make Vision and Language Models More G
rounded

Ramprasaath R. Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh
, Larry Heck, Dhruv Batra, Devi Parikh; Proceedings of the IEEE/CVF Internati
onal Conference on Computer Vision (ICCV), 2019, pp. 2591-2600

Many vision and language models suffer from poor visual grounding -- often falli
ng back on easy-to-learn language priors rather than basing their decisions on v
isual concepts in the image. In this work, we propose a generic approach called
Human Importance-aware Network Tuning (HINT) that effectively leverages human de
monstrations to improve visual grounding. HINT encourages deep networks to be se
nsitive to the same input regions as humans. Our approach optimizes the alignmen
t between human attention maps and gradient-based network importances -- ensurin
g that models learn not just to look at but rather rely on visual concepts that
humans found relevant for a task when making predictions. We apply HINT to Visua
l Question Answering and Image Captioning tasks, outperforming top approaches on
 splits that penalize over-reliance on language priors (VQA-CP and robust captio
ning) using human attention demonstrations for just 6% of the training data.
****

Align2Ground: Weakly Supervised Phrase Grounding Guided by Image-Caption Alignme
nt

Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, Ajay Div
akaran; Proceedings of the IEEE/CVF International Conference on Computer Vision
(ICCV), 2019, pp. 2601-2610

We address the problem of grounding free-form textual phrases by using weak supe
rvision from image-caption pairs. We propose a novel end-to-end model that uses

caption-to-image retrieval as a downstream task to guide the process of phrase l
ocalization. Our method, as a first step, infers the latent correspondences betw
een regions-of-interest (RoIs) and phrases in the caption and creates a discrimi
native image representation using these matched RoIs. In the subsequent step, th
is learned representation is aligned with the caption. Our key contribution lies
 in building this "caption-conditioned" image encoding, which tightly couples bo
th the tasks and allows the weak supervision to effectively guide visual groundi
ng. We provide extensive empirical and qualitative analysis to investigate the d
ifferent components of our proposed model and compare it with competitive baseli
nes. For phrase localization, we report an improvement of 4.9% and 1.3% (absolut
e) over the prior state-of-the-art on the VisualGenome and Flickr30k Entities da
tasets. We also report results that are at par with the state-of-the-art on the
downstream caption-to-image retrieval task on COCO and Flickr30k datasets.
****

Adaptive Reconstruction Network for Weakly Supervised Referring Expression Groun
ding

Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Dechao Meng, Qingming Hu
ang; Proceedings of the IEEE/CVF International Conference on Computer Vision (IC
CV), 2019, pp. 2611-2620

Weakly supervised referring expression grounding aims at localizing the referent
ial object in an image according to the linguistic query, where the mapping betw
een the referential object and query is unknown in the training stage. To addres
s this problem, we propose a novel end-to-end adaptive reconstruction network (A
RN). It builds the correspondence between image region proposal and query in an
adaptive manner: adaptive grounding and collaborative reconstruction. Specifical
ly, we first extract the subject, location and context features to represent the
 proposals and the query respectively. Then, we design the adaptive grounding mo
dule to compute the matching score between each proposal and query by a hierarch
ical attention model. Finally, based on attention score and proposal features, w
e reconstruct the input query with a collaborative loss of language reconstructi
on loss, adaptive reconstruction loss, and attribute classification loss. This a
daptive mechanism helps our model to alleviate the variance of different referri
ng expressions. Experiments on four large-scale datasets show ARN outperforms ex
isting state-of-the-art methods by a large margin. Qualitative results demonstra
te that the proposed ARN can better handle the situation where multiple objects
of a particular category situated together.
****

Hierarchy Parsing for Image Captioning

Ting Yao, Yingwei Pan, Yehao Li, Tao Mei; Proceedings of the IEEE/CVF Interna
tional Conference on Computer Vision (ICCV), 2019, pp. 2621-2629

It is always well believed that parsing an image into constituent visual pattern
s would be helpful for understanding and representing an image. Nevertheless, th
ere has not been evidence in support of the idea on describing an image with a n
atural-language utterance. In this paper, we introduce a new design to model a h
ierarchy from instance level (segmentation), region level (detection) to the who
le image to delve into a thorough image understanding for captioning. Specifical
ly, we present a HIerarchy Parsing (HIP) architecture that novelly integrates hi
erarchical structure into image encoder. Technically, an image decomposes into a
 set of regions and some of the regions are resolved into finer ones. Each regio
n then regresses to an instance, i.e., foreground of the region. Such process na
turally builds a hierarchal tree. A tree-structured Long Short-Term Memory (Tree
-LSTM) network is then employed to interpret the hierarchal structure and enhanc
e all the instance-level, region-level and image-level features. Our HIP is appe
aling in view that it is pluggable to any neural captioning models. Extensive ex
periments on COCO image captioning dataset demonstrate the superiority of HIP. M
ore remarkably, HIP plus a top-down attention-based LSTM decoder increases CIDEr
-D performance from 120.1% to 127.2% on COCO Karpathy test split. When further e
ndowing instance-level and region-level features from HIP with semantic relation
 learnt through Graph Convolutional Networks (GCN), CIDEr-D is boosted up to 130
.6%.

****
HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips
Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, Josef Sivic; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2630-2640
Learning text-video embeddings usually requires a dataset of video clips with manually provided captions. However, such datasets are expensive and time consuming to create and therefore difficult to obtain on a large scale. In this work, we propose instead to learn such embeddings from video data with readily available natural language annotations in the form of automatically transcribed narrations. The contributions of this work are three-fold. First, we introduce HowTo100M: a large-scale dataset of 136 million video clips sourced from 1.22M narrated instructional web videos depicting humans performing and describing over 23k different visual tasks. Our data collection procedure is fast, scalable and does not require any additional manual annotation. Second, we demonstrate that a text-video embedding trained on this data leads to state-of-the-art results for text-to-video retrieval and action localization on instructional video datasets such as YouCook2 or CrossTask. Finally, we show that this embedding transfers well to other domains: fine-tuning on generic Youtube videos (MSR-VTT dataset) and movies (LSMDC dataset) outperforms models trained on these datasets alone. Our dataset, code and models are publicly available.
****
Controllable Video Captioning With POS Sequence Guidance Based on Gated Fusion Network
Bairui Wang, Lin Ma, Wei Zhang, Wenhao Jiang, Jingwen Wang, Wei Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2641-2650
In this paper, we propose to guide the video caption generation with Part-of-Speech (POS) information, based on a gated fusion of multiple representations of input videos. We construct a novel gated fusion network, with one particularly designed cross-gating (CG) block, to effectively encode and fuse different types of representations, e.g., the motion and content features of an input video. One POS sequence generator relies on this fused representation to predict the global syntactic structure, which is thereafter leveraged to guide the video captioning generation and control the syntax of the generated sentence. Specifically, a gating strategy is proposed to dynamically and adaptively incorporate the global syntactic POS information into the decoder for generating each word. Experimental results on two benchmark datasets, namely MSR-VTT and MSVD, demonstrate that the proposed model can well exploit complementary information from multiple representations, resulting in improved performances. Moreover, the generated global POS information can well capture the global syntactic structure of the sentence, and thus be exploited to control the syntactic structure of the description. Such POS information not only boosts the video captioning performance but also improves the diversity of the generated captions. Our code is at: https://github.com/vsislab/Controllable_XGating.
****
Multi-View Stereo by Temporal Nonparametric Fusion
Yuxin Hou, Juho Kannala, Arno Solin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2651-2660
We propose a novel idea for depth estimation from multi-view image-pose pairs, where the model has capability to leverage information from previous latent-space encodings of the scene. This model uses pairs of images and poses, which are passed through an encoder-decoder model for disparity estimation. The novelty lies in soft-constraining the bottleneck layer by a nonparametric Gaussian process prior. We propose a pose-kernel structure that encourages similar poses to have resembling latent spaces. The flexibility of the Gaussian process (GP) prior provides adapting memory for fusing information from nearby views. We train the encoder-decoder and the GP hyperparameters jointly end-to-end. In addition to a batch method, we derive a lightweight estimation scheme that circumvents standard pi

tfalls in scaling Gaussian process inference, and demonstrate how our scheme can run in real-time on smart devices.
****

Floor-SP: Inverse CAD for Floorplans by Sequential Room-Wise Shortest Path
Jiacheng Chen, Chen Liu, Jiaye Wu, Yasutaka Furukawa; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2661-2670
This paper proposes a new approach for automated floorplan reconstruction from RGBD scans, a major milestone in indoor mapping research. The approach, dubbed Floor-SP, formulates a novel optimization problem, where room-wise coordinate descent sequentially solves shortest path problems to optimize the floorplan graph structure. The objective function consists of data terms guided by deep neural networks, consistency terms encouraging adjacent rooms to share corners and walls, and the model complexity term. The approach does not require corner/edge primitive extraction unlike most other methods. We have evaluated our system on production-quality RGBD scans of 527 apartments or houses, including many units with non-Manhattan structures. Qualitative and quantitative evaluations demonstrate a significant performance boost over the current state-of-the-art. Please refer to our project website http://jcchen.me/floor-sp/ for code and data.
****

Polarimetric Relative Pose Estimation
Zhaopeng Cui, Viktor Larsson, Marc Pollefeys; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2671-2680
In this paper we consider the problem of relative pose estimation from two images with per-pixel polarimetric information. Using these additional measurements we derive a simple minimal solver for the essential matrix which only requires two point correspondences. The polarization constraints allow us to pointwise recover the 3D surface normal up to a two-fold ambiguity for the diffuse reflection. Since this ambiguity exists per point, there is a combinatorial explosion of possibilities. However, since our solver only requires two point correspondences, we only need to consider 16 configurations when solving for the relative pose. Once the relative orientation is recovered, we show that it is trivial to resolve the ambiguity for the remaining points. For robustness, we also propose a joint optimization between the relative pose and the refractive index to handle the refractive distortion. In experiments, on both synthetic and real data, we demonstrate that by leveraging the additional information available from polarization cameras, we can improve over classical methods which only rely on the 2D-point locations to estimate the geometry. Finally, we demonstrate the practical applicability of our approach by integrating it into a state-of-the-art global Structure-from-Motion pipeline.
****

Closed-Form Optimal Two-View Triangulation Based on Angular Errors
Seong Hun Lee, Javier Civera; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2681-2689
In this paper, we study closed-form optimal solutions to two-view triangulation with known internal calibration and pose. By formulating the triangulation problem as L-1 and L-infinity minimization of angular reprojection errors, we derive the exact closed-form solutions that guarantee global optimality under respective cost functions. To the best of our knowledge, we are the first to present such solutions. Since the angular error is rotationally invariant, our solutions can be applied for any type of central cameras, be it perspective, fisheye or omnidirectional. Our methods also require significantly less computation than the existing optimal methods. Experimental results on synthetic and real datasets validate our theoretical derivations.
****

Pix2Vox: Context-Aware 3D Reconstruction From Single and Multi-View Images
Haozhe Xie, Hongxun Yao, Xiaoshuai Sun, Shangchen Zhou, Shengping Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2690-2698
Recovering the 3D representation of an object from single-view or multi-view RGB images by deep neural networks has attracted increasing attention in the past f

ew years. Several mainstream works (e.g., 3D-R2N2) use recurrent neural networks (RNNs) to fuse multiple feature maps extracted from input images sequentially. However, when given the same set of input images with different orders, RNN-based approaches are unable to produce consistent reconstruction results. Moreover, due to long-term memory loss, RNNs cannot fully exploit input images to refine reconstruction results. To solve these problems, we propose a novel framework for single-view and multi-view 3D reconstruction, named Pix2Vox. By using a well-designed encoder-decoder, it generates a coarse 3D volume from each input image. Then, a context-aware fusion module is introduced to adaptively select high-quality reconstructions for each part (e.g., table legs) from different coarse 3D volumes to obtain a fused 3D volume. Finally, a refiner further refines the fused 3D volume to generate the final output. Experimental results on the ShapeNet and Pix3D benchmarks indicate that the proposed Pix2Vox outperforms state-of-the-arts by a large margin. Furthermore, the proposed method is 24 times faster than 3D-R2N2 in terms of backward inference time. The experiments on ShapeNet unseen 3D categories have shown the superior generalization abilities of our method.
****

Unsupervised Robust Disentangling of Latent Characteristics for Image Synthesis
Patrick Esser, Johannes Haux, Bjorn Ommer; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2699-2709
Deep generative models come with the promise to learn an explainable representation for visual objects that allows image sampling, synthesis, and selective modification. The main challenge is to learn to properly model the independent latent characteristics of an object, especially its appearance and pose. We present a novel approach that learns disentangled representations of these characteristics and explains them individually. Training requires only pairs of images depicting the same object appearance, but no pose annotations. We propose an additional classifier that estimates the minimal amount of regularization required to enforce disentanglement. Thus both representations together can completely explain an image while being independent of each other. Previous methods based on adversarial approaches fail to enforce this independence, while methods based on variational approaches lead to uninformative representations. In experiments on diverse object categories, the approach successfully recombines pose and appearance to reconstruct and retarget novel synthesized images. We achieve significant improvements over state-of-the-art methods which utilize the same level of supervision, and reach performances comparable to those of pose-supervised approaches. However, we can handle the vast body of articulated object classes for which no pose models/annotations are available.
****

SROBB: Targeted Perceptual Loss for Single Image Super-Resolution
Mohammad Saeed Rad, Behzad Bozorgtabar, Urs-Viktor Marti, Max Basler, Hazim Kemal Ekenel, Jean-Philippe Thiran; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2710-2719
By benefiting from perceptual losses, recent studies have improved significantly the performance of the super-resolution task, where a high-resolution image is resolved from its low-resolution counterpart. Although such objective functions generate near-photorealistic results, their capability is limited, since they estimate the reconstruction error for an entire image in the same way, without considering any semantic information. In this paper, we propose a novel method to benefit from perceptual loss in a more objective way. We optimize a deep network-based decoder with a targeted objective function that penalizes images at different semantic levels using the corresponding terms. In particular, the proposed method leverages our proposed OBB (Object, Background and Boundary) labels, generated from segmentation labels, to estimate a suitable perceptual loss for boundaries, while considering texture similarity for backgrounds. We show that our proposed approach results in more realistic textures and sharper edges, and outperforms other state-of-the-art algorithms in terms of both qualitative results on standard benchmarks and results of extensive user studies.
****

An Internal Learning Approach to Video Inpainting

Haotian Zhang, Long Mai, Ning Xu, Zhaowen Wang, John Collomosse, Hailin Jin ; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) , 2019, pp. 2720-2729
We propose a novel video inpainting algorithm that simultaneously hallucinates m issing appearance and motion (optical flow) information, building upon the recen t 'Deep Image Prior' (DIP) that exploits convolutional network architectures to enforce plausible texture in static images. In extending DIP to video we make tw o important contributions. First, we show that coherent video inpainting is poss ible without a priori training. We take a generative approach to inpainting base d on internal (within-video) learning without reliance upon an external corpus o f visual data to train a one-size-fits-all model for the large space of general videos. Second, we show that such a framework can jointly generate both appearan ce and flow, whilst exploiting these complementary modalities to ensure mutual c onsistency. We show that leveraging appearance statistics specific to each video achieves visually plausible results whilst handling the challenging problem of long-term consistency.
****

Deep CG2Real: Synthetic-to-Real Translation via Image Disentanglement
Sai Bi, Kalyan Sunkavalli, Federico Perazzi, Eli Shechtman, Vladimir G. Kim, Ravi Ramamoorthi; Proceedings of the IEEE/CVF International Conference on Comp uter Vision (ICCV), 2019, pp. 2730-2739
We present a method to improve the visual realism of low-quality, synthetic imag es, e.g. OpenGL renderings. Training an unpaired synthetic-to-real translation n etwork in image space is severely under-constrained and produces visible artifac ts. Instead, we propose a semi-supervised approach that operates on the disentan gled shading and albedo layers of the image. Our two-stage pipeline first learns to predict accurate shading in a supervised fashion using physically-based rend erings as targets, and further increases the realism of the textures and shading with an improved CycleGAN network. Extensive evaluations on the SUNCG indoor sc ene dataset demonstrate that our approach yields more realistic images compared to other state-of-the-art approaches. Furthermore, networks trained on our gener ated "real" images predict more accurate depth and normals than domain adaptatio n approaches, suggesting that improving the visual realism of the images can be more effective than imposing task-specific losses.
****

Adversarial Defense via Learning to Generate Diverse Attacks
Yunseok Jang, Tianchen Zhao, Seunghoon Hong, Honglak Lee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2740-2749
With the remarkable success of deep learning, Deep Neural Networks (DNNs) have b een applied as dominant tools to various machine learning domains. Despite this success, however, it has been found that DNNs are surprisingly vulnerable to mal icious attacks; adding a small, perceptually indistinguishable perturbations to the data can easily degrade classification performance. Adversarial training is an effective defense strategy to train a robust classifier. In this work, we pro pose to utilize the generator to learn how to create adversarial examples. Unlik e the existing approaches that create a one-shot perturbation by a deterministic generator, we propose a recursive and stochastic generator that produces much s tronger and diverse perturbations that comprehensively reveal the vulnerability of the target classifier. Our experiment results on MNIST and CIFAR-10 datasets show that the classifier adversarially trained with our method yields more robus t performance over various white-box and black-box attacks.
****

Image Generation From Small Datasets via Batch Statistics Adaptation
Atsuhiro Noguchi, Tatsuya Harada; Proceedings of the IEEE/CVF International Con ference on Computer Vision (ICCV), 2019, pp. 2750-2758
Thanks to the recent development of deep generative models, it is becoming possi ble to generate high-quality images with both fidelity and diversity. However, t he training of such generative models requires a large dataset. To reduce the am ount of data required, we propose a new method for transferring prior knowledge of the pre-trained generator, which is trained with a large dataset, to a small

dataset in a different domain. Using such prior knowledge, the model can generate images leveraging some common sense that cannot be acquired from a small dataset. In this work, we propose a novel method focusing on the parameters for batch statistics, scale and shift, of the hidden layers in the generator. By training only these parameters in a supervised manner, we achieved stable training of the generator, and our method can generate higher quality images compared to previous methods without collapsing, even when the dataset is small (100). Our results show that the diversity of the filters acquired in the pre-trained generator is important for the performance on the target domain. Our method makes it possible to add a new class or domain to a pre-trained generator without disturbing the performance on the original domain. Code is available at github.com/nogu-atsu/small-dataset-image-generation

****

Lifelong GAN: Continual Learning for Conditional Image Generation

Mengyao Zhai, Lei Chen, Frederick Tung, Jiawei He, Megha Nawhal, Greg Mori; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2759-2768

Lifelong learning is challenging for deep neural networks due to their susceptibility to catastrophic forgetting. Catastrophic forgetting occurs when a trained network is not able to maintain its ability to accomplish previously learned tasks when it is trained to perform new tasks. We study the problem of lifelong learning for generative models, extending a trained network to new conditional generation tasks without forgetting previous tasks, while assuming access to the training data for the current task only. In contrast to state-of-the-art memory replay based approaches which are limited to label-conditioned image generation tasks, a more generic framework for continual learning of generative models under different conditional image generation settings is proposed in this paper. Lifelong GAN employs knowledge distillation to transfer learned knowledge from previous networks to the new network. This makes it possible to perform image-conditioned generation tasks in a lifelong learning setting. We validate Lifelong GAN for both image-conditioned and label-conditioned generation tasks, and provide qualitative and quantitative results to show the generality and effectiveness of our method.

****

Bayesian Relational Memory for Semantic Visual Navigation

Yi Wu, Yuxin Wu, Aviv Tamar, Stuart Russell, Georgia Gkioxari, Yuandong Tian; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2769-2779

We introduce a new memory architecture, Bayesian Relational Memory (BRM), to improve the generalization ability for semantic visual navigation agents in unseen environments, where an agent is given a semantic target to navigate towards. BRM takes the form of a probabilistic relation graph over semantic entities (e.g., room types), which allows (1) capturing the layout prior from training environments, i.e., prior knowledge, (2) estimating posterior layout at test time, i.e., memory update, and (3) efficient planning for navigation, altogether. We develop a BRM agent consisting of a BRM module for producing sub-goals and a goal-conditioned locomotion module for control. When testing in unseen environments, the BRM agent outperforms baselines that do not explicitly utilize the probabilistic relational memory structure.

****

Mono-SF: Multi-View Geometry Meets Single-View Depth for Monocular Scene Flow Estimation of Dynamic Traffic Scenes

Fabian Brickwedde, Steffen Abraham, Rudolf Mester; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2780-2790

Existing 3D scene flow estimation methods provide the 3D geometry and 3D motion of a scene and gain a lot of interest, for example in the context of autonomous driving. These methods are traditionally based on a temporal series of stereo images. In this paper, we propose a novel monocular 3D scene flow estimation method, called Mono-SF. Mono-SF jointly estimates the 3D structure and motion of the scene by combining multi-view geometry and single-view depth information. Mono-S

F considers that the scene flow should be consistent in terms of warping the ref erence image in the consecutive image based on the principles of multi-view geom etry. For integrating single-view depth in a statistical manner, a convolutional neural network, called ProbDepthNet, is proposed. ProbDepthNet estimates pixel-wise depth distributions from a single image rather than single depth values. Ad ditionally, as part of ProbDepthNet, a novel recalibration technique for regress ion problems is proposed to ensure well-calibrated distributions. Our experiment s show that Mono-SF outperforms state-of-the-art monocular baselines and ablatio n studies support the Mono-SF approach and ProbDepthNet design.
****

Prior Guided Dropout for Robust Visual Localization in Dynamic Environments
Zhaoyang Huang, Yan Xu, Jianping Shi, Xiaowei Zhou, Hujun Bao, Guofeng Zhan g; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2791-2800
Camera localization from monocular images has been a long-standing problem, but its robustness in dynamic environments is still not adequately addressed. Compar ed with classic geometric approaches, modern CNN-based methods (e.g. PoseNet) ha ve manifested the reliability against illumination or viewpoint variations, but they still have the following limitations. First, foreground moving objects are not explicitly handled, which results in poor performance and instability in dyn amic environments. Second, the output for each image is a point estimate without uncertainty quantification. In this paper, we propose a framework which can be generally applied to existing CNN-based pose regressors to improve their robustn ess in dynamic environments. The key idea is a prior guided dropout module coupl ed with a self-attention module which can guide CNNs to ignore foreground object s during both training and inference. Additionally, the dropout module enables t he pose regressor to output multiple hypotheses from which the uncertainty of po se estimates can be quantified and leveraged in the following uncertainty-aware pose-graph optimization to improve the robustness further. We achieve an average accuracy of 9.98m/3.63deg on RobotCar dataset, which outperforms the state-of-t he-art method by 62.97%/47.08%. The source code of our implementation is availab le at https://github.com/zju3dv/RVL-dynamic.
****

Drive&Act: A Multi-Modal Dataset for Fine-Grained Driver Behavior Recognition in Autonomous Vehicles
Manuel Martin, Alina Roitberg, Monica Haurilet, Matthias Horne, Simon Reiss, Michael Voit, Rainer Stiefelhagen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2801-2810
We introduce the novel domain-specific Drive&Act benchmark for fine-grained cate gorization of driver behavior. Our dataset features twelve hours and over 9.6 mi llion frames of people engaged in distractive activities during both, manual and automated driving. We capture color, infrared, depth and 3D body pose informati on from six views and densely label the videos with a hierarchical annotation sc heme, resulting in 83 categories. The key challenges of our dataset are: (1) rec ognition of fine-grained behavior inside the vehicle cabin; (2) multi-modal acti vity recognition, focusing on diverse data streams; and (3) a cross view recogni tion benchmark, where a model handles data from an unfamiliar domain, as sensor type and placement in the cabin can change between vehicles. Finally, we provide challenging benchmarks by adopting prominent methods for video- and body pose-b ased action recognition.
****

Depth Completion From Sparse LiDAR Data With Depth-Normal Constraints
Yan Xu, Xinge Zhu, Jianping Shi, Guofeng Zhang, Hujun Bao, Hongsheng Li; Pr oceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 20 19, pp. 2811-2820
Depth completion aims to recover dense depth maps from sparse depth measurement s. It is of increasing importance for autonomous driving and draws increasing att ention from the vision community. Most of the current competitive methods direct ly train a network to learn a mapping from sparse depth inputs to dense depth ma ps, which has difficulties in utilizing the 3D geometric constraints and handlin

g the practical sensor noises. In this paper, to regularize the depth completion and improve the robustness against noise, we propose a unified CNN framework that 1) models the geometric constraints between depth and surface normal in a diffusion module and 2) predicts the confidence of sparse LiDAR measurements to mitigate the impact of noise. Specifically, our encoder-decoder backbone predicts the surface normal, coarse depth and confidence of LiDAR inputs simultaneously, which are subsequently inputted into our diffusion refinement module to obtain the final completion results. Extensive experiments on KITTI depth completion dataset and NYU-Depth-V2 dataset demonstrate that our method achieves state-of-the-art performance. Further ablation study and analysis give more insights into the proposed components and demonstrate the generalization capability and stability of our model.

****

PRECOG: PREdiction Conditioned on Goals in Visual Multi-Agent Settings

Nicholas Rhinehart, Rowan McAllister, Kris Kitani, Sergey Levine; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2821-2830

For autonomous vehicles (AVs) to behave appropriately on roads populated by human-driven vehicles, they must be able to reason about the uncertain intentions and decisions of other drivers from rich perceptual information. Towards these capabilities, we present a probabilistic forecasting model of future interactions between a variable number of agents. We perform both standard forecasting and the novel task of conditional forecasting, which reasons about how all agents will likely respond to the goal of a controlled agent (here, the AV). We train models on real and simulated data to forecast vehicle trajectories given past positions and LIDAR. Our evaluation shows that our model is substantially more accurate in multi-agent driving scenarios compared to existing state-of-the-art. Beyond its general ability to perform conditional forecasting queries, we show that our model's predictions of all agents improve when conditioned on knowledge of the AV's goal, further illustrating its capability to model agent interactions.

****

LPD-Net: 3D Point Cloud Learning for Large-Scale Place Recognition and Environment Analysis

Zhe Liu, Shunbo Zhou, Chuanzhe Suo, Peng Yin, Wen Chen, Hesheng Wang, Haoang Li, Yun-Hui Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2831-2840

Point cloud based place recognition is still an open issue due to the difficulty in extracting local features from the raw 3D point cloud and generating the global descriptor, and it's even harder in the large-scale dynamic environments. In this paper, we develop a novel deep neural network, named LPD-Net (Large-scale Place Description Network), which can extract discriminative and generalizable global descriptors from the raw 3D point cloud. Two modules, the adaptive local feature extraction module and the graph-based neighborhood aggregation module, are proposed, which contribute to extract the local structures and reveal the spatial distribution of local features in the large-scale point cloud, with an end-to-end manner. We implement the proposed global descriptor in solving point cloud based retrieval tasks to achieve the large-scale place recognition. Comparison results show that our LPD-Net is much better than PointNetVLAD and reaches the state-of-the-art. We also compare our LPD-Net with the vision-based solutions to show the robustness of our approach to different weather and light conditions.

****

Local Supports Global: Deep Camera Relocalization With Sequence Enhancement

Fei Xue, Xin Wang, Zike Yan, Qiuyuan Wang, Junqiu Wang, Hongbin Zha; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2841-2850

We propose to leverage the local information in a image sequence to support global camera relocalization. In contrast to previous methods that regress global poses from single images, we exploit the spatial-temporal consistency in sequential images to alleviate uncertainty due to visual ambiguities by incorporating a visual odometry (VO) component. Specifically, we introduce two effective steps ca

lled content-augmented pose estimation and motion-based refinement. The content-augmentation step focuses on alleviating the uncertainty of pose estimation by augmenting the observation based on the co-visibility in local maps built by the VO stream. Besides, the motion-based refinement is formulated as a pose graph, where the camera poses are further optimized by adopting relative poses provided by the VO component as additional motion constraints. Thus, the global consistency can be guaranteed. Experiments on the public indoor 7-Scenes and outdoor Oxford RobotCar benchmark datasets demonstrate that benefited from local information inherent in the sequence, our approach outperforms state-of-the-art methods, especially in some challenging cases, e.g., insufficient texture, highly repetitive textures, similar appearances, and over-exposure.
****

Sequential Adversarial Learning for Self-Supervised Deep Visual Odometry
Shunkai Li, Fei Xue, Xin Wang, Zike Yan, Hongbin Zha; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2851-2860
We propose a self-supervised learning framework for visual odometry (VO) that incorporates correlation of consecutive frames and takes advantage of adversarial learning. Previous methods tackle self-supervised VO as a local structure from motion (SfM) problem that recovers depth from single image and relative poses from image pairs by minimizing photometric loss between warped and captured images. As single-view depth estimation is an ill-posed problem, and photometric loss is incapable of discriminating distortion artifacts of warped images, the estimated depth is vague and pose is inaccurate. In contrast to previous methods, our framework learns a compact representation of frame-to-frame correlation, which is updated by incorporating sequential information. The updated representation is used for depth estimation. Besides, we tackle VO as a self-supervised image generation task and take advantage of Generative Adversarial Networks (GAN). The generator learns to estimate depth and pose to generate a warped target image. The discriminator evaluates the quality of generated image with high-level structural perception that overcomes the problem of pixel-wise loss in previous methods. Experiments on KITTI and Cityscapes datasets show that our method obtains more accurate depth with details preserved and predicted pose outperforms state-of-the-art self-supervised methods significantly.
****

TextPlace: Visual Place Recognition and Topological Localization Through Reading Scene Texts
Ziyang Hong, Yvan Petillot, David Lane, Yishu Miao, Sen Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2861-2870
Visual place recognition is a fundamental problem for many vision based applications. Sparse feature and deep learning based methods have been successful and dominant over the decade. However, most of them do not explicitly leverage high-level semantic information to deal with challenging scenarios where they may fail. This paper proposes a novel visual place recognition algorithm, termed TextPlace, based on scene texts in the wild. Since scene texts are high-level information invariant to illumination changes and very distinct for different places when considering spatial correlation, it is beneficial for visual place recognition tasks under extreme appearance changes and perceptual aliasing. It also takes spatial-temporal dependence between scene texts into account for topological localization. Extensive experiments show that TextPlace achieves state-of-the-art performance, verifying the effectiveness of using high-level scene texts for robust visual place recognition in urban areas.
****

CamNet: Coarse-to-Fine Retrieval for Camera Re-Localization
Mingyu Ding, Zhe Wang, Jiankai Sun, Jianping Shi, Ping Luo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2871-2880
Camera re-localization is an important but challenging task in applications like robotics and autonomous driving. Recently, retrieval-based methods have been considered as a promising direction as they can be easily generalized to novel sce

nes. Despite significant progress has been made, we observe that the performance bottleneck of previous methods actually lies in the retrieval module. These methods use the same features for both retrieval and relative pose regression tasks which have potential conflicts in learning. To this end, here we present a coarse-to-fine retrieval-based deep learning framework, which includes three steps, i.e., image-based coarse retrieval, pose-based fine retrieval and precise relative pose regression. With our carefully designed retrieval module, the relative pose regression task can be surprisingly simpler. We design novel retrieval losses with batch hard sampling criterion and two-stage retrieval to locate samples that adapt to the relative pose regression task. Extensive experiments show that our model (CamNet) outperforms the state-of-the-art methods by a large margin on both indoor and outdoor datasets.

****

Situational Fusion of Visual Representation for Visual Navigation

William B. Shen, Danfei Xu, Yuke Zhu, Leonidas J. Guibas, Li Fei-Fei, Silvio Savarese; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2881-2890

A complex visual navigation task puts an agent in different situations which call for a diverse range of visual perception abilities. For example, to "go to the nearest chair", the agent might need to identify a chair in a living room using semantics, follow along a hallway using vanishing point cues, and avoid obstacles using depth. Therefore, utilizing the appropriate visual perception abilities based on a situational understanding of the visual environment can empower these navigation models in unseen visual environments. We propose to train an agent to fuse a large set of visual representations that correspond to diverse visual perception abilities. To fully utilize each representation, we develop an action-level representation fusion scheme, which predicts an action candidate from each representation and adaptively consolidate these action candidates into the final action. Furthermore, we employ a data-driven inter-task affinity regularization to reduce redundancies and improve generalization. Our approach leads to a significantly improved performance in novel environments over ImageNet-pretrained baseline and other fusion methods.

****

Learning Aberrance Repressed Correlation Filters for Real-Time UAV Tracking

Ziyuan Huang, Changhong Fu, Yiming Li, Fuling Lin, Peng Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2891-2900

Traditional framework of discriminative correlation filters (DCF) is often subject to undesired boundary effects. Several approaches to enlarge search regions have been already proposed in the past years to make up for this shortcoming. However, with excessive background information, more background noises are also introduced and the discriminative filter is prone to learn from the ambiance rather than the object. This situation, along with appearance changes of objects caused by full/partial occlusion, illumination variation, and other reasons has made it more likely to have aberrances in the detection process, which could substantially degrade the credibility of its result. Therefore, in this work, a novel approach to repress the aberrances happening during the detection process is proposed, i.e., aberrance repressed correlation filter (ARCF). By enforcing restriction to the rate of alteration in response maps generated in the detection phase, the ARCF tracker can evidently suppress aberrances and is thus more robust and accurate to track objects. Considerable experiments are conducted on different UAV datasets to perform object tracking from an aerial view, i.e., UAV123, UAVDT, and DTB70, with 243 challenging image sequences containing over 90K frames to verify the performance of the ARCF tracker and it has proven itself to have outperformed other 20 state-of-the-art trackers based on DCF and deep-based frameworks with sufficient speed for real-time applications.

****

6-DOF GraspNet: Variational Grasp Generation for Object Manipulation

Arsalan Mousavian, Clemens Eppner, Dieter Fox; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2901-2910

Generating grasp poses is a crucial component for any robot object manipulation task. In this work, we formulate the problem of grasp generation as sampling a set of grasps using a variational autoencoder and assess and refine the sampled grasps using a grasp evaluator model. Both Grasp Sampler and Grasp Refinement networks take 3D point clouds observed by a depth camera as input. We evaluate our approach in simulation and real-world robot experiments. Our approach achieves 88% success rate on various commonly used objects with diverse appearances, scales, and weights. Our model is trained purely in simulation and works in the real-world without any extra steps.
****

DAGMapper: Learning to Map by Discovering Lane Topology
Namdar Homayounfar, Wei-Chiu Ma, Justin Liang, Xinyu Wu, Jack Fan, Raquel Urtasun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2911-2920
One of the fundamental challenges to scale self-driving is being able to create accurate high definition maps (HD maps) with low cost. Current attempts to automate this pro- cess typically focus on simple scenarios, estimate independent maps per frame or do not have the level of precision required by modern self driving vehicles. In contrast, in this paper we focus on drawing the lane boundaries of complex highways with many lanes that contain topology changes due to forks and merges. Towards this goal, we formulate the problem as inference in a directed acyclic graphical model (DAG), where the nodes of the graph encode geo- metric and topological properties of the local regions of the lane boundaries. Since we do not know a priori the topology of the lanes, we also infer the DAG topology (i.e., nodes and edges) for each region. We demonstrate the effectiveness of our approach on two major North American Highways in two different states and show high precision and recall as well as 89% correct topology.
****

3D-LaneNet: End-to-End 3D Multiple Lane Detection
Noa Garnett, Rafi Cohen, Tomer Pe'er, Roee Lahav, Dan Levi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2921-2930
We introduce a network that directly predicts the 3D layout of lanes in a road scene from a single image. This work marks a first attempt to address this task with on-board sensing without assuming a known constant lane width or relying on pre-mapped environments. Our network architecture, 3D-LaneNet, applies two new concepts: intra-network inverse-perspective mapping (IPM) and anchor-based lane representation. The intra-network IPM projection facilitates a dual-representation information flow in both regular image-view and top-view. An anchor-per-column output representation enables our end-to-end approach which replaces common heuristics such as clustering and outlier rejection, casting lane estimation as an object detection problem. In addition, our approach explicitly handles complex situations such as lane merges and splits. Results are shown on two new 3D lane datasets, a synthetic and a real one. For comparison with existing methods, we test our approach on the image-only tuSimple lane detection benchmark, achieving performance competitive with state-of-the-art.
****

Once a MAN: Towards Multi-Target Attack via Learning Multi-Target Adversarial Network Once
Jiangfan Han, Xiaoyi Dong, Ruimao Zhang, Dongdong Chen, Weiming Zhang, Nenghai Yu, Ping Luo, Xiaogang Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5158-5167
Modern deep neural networks are often vulnerable to adversarial samples. Based on the first optimization-based attacking method, many following methods are proposed to improve the attacking performance and speed. Recently, generation-based methods have received much attention since they directly use feed-forward networks to generate the adversarial samples, which avoid the time-consuming iterative attacking procedure in optimization-based and gradient-based methods. However, current generation-based methods are only able to attack one specific target (category) within one model, thus making them not applicable to real classification

systems that often have hundreds/thousands of categories. In this paper, we propose the first Multi-target Adversarial Network (MAN), which can generate multi-target adversarial samples with a single model. By incorporating the specified category information into the intermediate features, it can attack any category of the target classification model during runtime. Experiments show that the proposed MAN can produce stronger attack results and also have better transferability than previous state-of-the-art methods in both multi-target attack task and single-target attack task. We further use the adversarial samples generated by our MAN to improve the robustness of the classification model. It can also achieve better classification accuracy than other methods when attacked by various methods.

****

Attention Bridging Network for Knowledge Transfer

Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, Yun Fu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5198-5207

The attention of a deep neural network obtained by back-propagating gradients can effectively explain the decision of the network. They can further be used to explicitly access to the network response to a specific pattern. Considering objects of the same category but from different domains share similar visual patterns, we propose to treat the network attention as a bridge to connect objects across domains. In this paper, we use knowledge from the source domain to guide the network's response to categories shared with the target domain. With weights sharing and domain adversary training, this knowledge can be successfully transferred by regularizing the network's response to the same category in the target domain. Specifically, we transfer the foreground prior from a simple single-label dataset to another complex multi-label dataset, leading to improvement of attention maps. Experiments about the weakly-supervised semantic segmentation task show the effectiveness of our method. Besides, we further explore and validate that the proposed method is able to improve the generalization ability of a classification network in domain adaptation and domain generalization settings.

****

Recover and Identify: A Generative Dual Model for Cross-Resolution Person Re-Identification

Yu-Jhe Li, Yun-Chun Chen, Yen-Yu Lin, Xiaofei Du, Yu-Chiang Frank Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8090-8099

Person re-identification (re-ID) aims at matching images of the same identity across camera views. Due to varying distances between cameras and persons of interest, resolution mismatch can be expected, which would degrade person re-ID performance in real-world scenarios. To overcome this problem, we propose a novel generative adversarial network to address cross-resolution person re-ID, allowing query images with varying resolutions. By advancing adversarial learning techniques, our proposed model learns resolution-invariant image representations while being able to recover the missing details in low-resolution input images. The resulting features can be jointly applied for improving person re-ID performance due to preserving resolution invariance and recovering re-ID oriented discriminative details. Our experiments on five benchmark datasets confirm the effectiveness of our approach and its superiority over the state-of-the-art methods, especially when the input resolutions are unseen during training.

****

Aggregation via Separation: Boosting Facial Landmark Detector With Semi-Supervised Style Translation

Shengju Qian, Keqiang Sun, Wayne Wu, Chen Qian, Jiaya Jia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10153-10163

Facial landmark detection, or face alignment, is a fundamental task that has been extensively studied. In this paper, we investigate a new perspective of facial landmark detection and demonstrate it leads to further notable improvement. Given that any face images can be factored into space of style that captures lighting, texture and image environment, and a style-invariant structure space, our ke

y idea is to leverage disentangled style and shape space of each individual to a
ugment existing structures via style translation. With these augmented synthetic
 samples, our semi-supervised model surprisingly outperforms the fully-supervise
d one by a large margin. Extensive experiments verify the effectiveness of our i
dea with state-of-the-art results on WFLW, 300W, COFW, and AFLW datasets. Our pr
oposed structure is general and could be assembled into any face alignment frame
works. The code is made publicly available at https://github.com/thesouthfrog/st
ylealign.
****