Finding Task-Relevant Features for Few-Shot Learning by Category Traversal
Hongyang Li, David Eigen, Samuel Dodge, Matthew Zeiler, Xiaogang Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1-10
Few-shot learning is an important area of research. Conceptually, humans are readily able to understand new concepts given just a few examples, while in more pragmatic terms, limited-example training situations are common practice. Recent effective approaches to few-shot learning employ a metric-learning framework to learn a feature similarity comparison between a query (test) example, and the few support (training) examples. However, these approaches treat each support class independently from one another, never looking at the entire task as a whole. Because of this, they are constrained to use a single set of features for all possible test-time tasks, which hinders the ability to distinguish the most relevant dimensions for the task at hand. In this work, we introduce a Category Traversal Module that can be inserted as a plug-and-play module into most metric-learning based few-shot learners. This component traverses across the entire support set at once, identifying task-relevant features based on both intra-class commonality and inter-class uniqueness in the feature space. Incorporating our module improves performance considerably (5%-10% relative) over baseline systems on both miniImageNet and tieredImageNet benchmarks, with overall performance competitive with the most recent state-of-the-art systems.
****
Edge-Labeling Graph Neural Network for Few-Shot Learning
Jongmin Kim, Taesup Kim, Sungwoong Kim, Chang D. Yoo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11-20
In this paper, we propose a novel edge-labeling graph neural network (EGNN), which adapts a deep neural network on the edge-labeling graph, for few-shot learning. The previous graph neural network (GNN) approaches in few-shot learning have been based on the node-labeling framework, which implicitly models the intra-cluster similarity and the inter-cluster dissimilarity. In contrast, the proposed EGNN learns to predict the edge-labels rather than the node-labels on the graph that enables the evolution of an explicit clustering by iteratively updating the edge-labels with direct exploitation of both intra-cluster similarity and the inter-cluster dissimilarity. It is also well suited for performing on various numbers of classes without retraining, and can be easily extended to perform a transductive inference. The parameters of the EGNN are learned by episodic training with an edge-labeling loss to obtain a well-generalizable model for unseen low-data problem. On both of the supervised and semi-supervised few-shot image classification tasks with two benchmark datasets, the proposed EGNN significantly improves the performances over the existing GNNs.
****
Generating Classification Weights With GNN Denoising Autoencoders for Few-Shot Learning
Spyros Gidaris, Nikos Komodakis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 21-30
Given an initial recognition model already trained on a set of base classes, the goal of this work is to develop a meta-model for few-shot learning. The meta-model, given as input some novel classes with few training examples per class, must properly adapt the existing recognition model into a new model that can correctly classify in a unified way both the novel and the base classes. To accomplish this goal it must learn to output the appropriate classification weight vectors for those two types of classes. To build our meta-model we make use of two main innovations: we propose the use of a Denoising Autoencoder network (DAE) that (during training) takes as input a set of classification weights corrupted with Gaussian noise and learns to reconstruct the target-discriminative classification weights. In this case, the injected noise on the classification weights serves the role of regularizing the weight generating meta-model. Furthermore, in order to capture the co-dependencies between different classes in a given task instance of our meta-model, we propose to implement the DAE model as a Graph Neural Ne

twork (GNN). In order to verify the efficacy of our approach, we extensively eva luate it on ImageNet based few-shot benchmarks and we report state-of-the-art re sults.
****

Kervolutional Neural Networks
Chen Wang, Jianfei Yang, Lihua Xie, Junsong Yuan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 31-40
Convolutional neural networks (CNNs) have enabled the state-of-the-art performan ce in many computer vision tasks. However, little effort has been devoted to est ablishing convolution in non-linear space. Existing works mainly leverage on the activation layers, which can only provide point-wise non-linearity. To solve th is problem, a new operation, kervolution (kernel convolution), is introduced to approximate complex behaviors of human perception systems leveraging on the kern el trick. It generalizes convolution, enhances the model capacity, and captures higher order interactions of features, via patch-wise kernel functions, but with out introducing additional parameters. Extensive experiments show that kervoluti onal neural networks (KNN) achieve higher accuracy and faster convergence than b aseline CNN.
****

Why ReLU Networks Yield High-Confidence Predictions Far Away From the Training D ata and How to Mitigate the Problem
Matthias Hein, Maksym Andriushchenko, Julian Bitterwolf; Proceedings of the IE EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4 1-50
Classifiers used in the wild, in particular for safety-critical systems, should not only have good generalization properties but also should know when they don' t know, in particular make low confidence predictions far away from the training data. We show that ReLU type neural networks which yield a piecewise linear cla ssifier function fail in this regard as they produce almost always high confiden ce predictions far away from the training data. For bounded domains like images we propose a new robust optimization technique similar to adversarial training w hich enforces low confidence predictions far away from the training data. We sho w that this technique is surprisingly effective in reducing the confidence of pr edictions far away from the training data while maintaining high confidence pred ictions and test error on the original classification task compared to standard training.
****

On the Structural Sensitivity of Deep Convolutional Networks to the Directions o f Fourier Basis Functions
Yusuke Tsuzuku, Issei Sato; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 51-60
Data-agnostic quasi-imperceptible perturbations on inputs are known to degrade r ecognition accuracy of deep convolutional networks severely. This phenomenon is considered to be a potential security issue. Moreover, some results on statistic al generalization guarantees indicate that the phenomena can be a key to improve the networks' generalization. However, the characteristics of the shared direct ions of such harmful perturbations remain unknown. Our primal finding is that co nvolutional networks are sensitive to the directions of Fourier basis functions. We derived the property by specializing a hypothesis of the cause of the sensit ivity, known as the linearity of neural networks, to convolutional networks and empirically validated it. As a byproduct of the analysis, we propose an algorith m to create shift-invariant universal adversarial perturbations available in bla ck-box settings.
****

Neural Rejuvenation: Improving Deep Network Training by Enhancing Computational Resource Utilization
Siyuan Qiao, Zhe Lin, Jianming Zhang, Alan L. Yuille; Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 61-71
In this paper, we study the problem of improving computational resource utilizat

ion of neural networks. Deep neural networks are usually over-parameterized for their tasks in order to achieve good performances, thus are likely to have under utilized computational resources. This observation motivates a lot of research t opics, e.g. network pruning, architecture search, etc. As models with higher com putational costs (e.g. more parameters or more computations) usually have better performances, we study the problem of improving the resource utilization of neu ral networks so that their potentials can be further realized. To this end, we p ropose a novel optimization method named Neural Rejuvenation. As its name sugges ts, our method detects dead neurons and computes resource utilization in real ti me, rejuvenates dead neurons by resource reallocation and reinitialization, and trains them with new training schemes. By simply replacing standard optimizers w ith Neural Rejuvenation, we are able to improve the performances of neural netwo rks by a very large margin while using similar training efforts and maintaining their original resource usages. The code is available here: https://github.com/j oe-siyuan-qiao/NeuralRejuvenation-CVPR19

****

Hardness-Aware Deep Metric Learning

Wenzhao Zheng, Zhaodong Chen, Jiwen Lu, Jie Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 72-81

This paper presents a hardness-aware deep metric learning (HDML) framework. Most previous deep metric learning methods employ the hard negative mining strategy to alleviate the lack of informative samples for training. However, this mining strategy only utilizes a subset of training data, which may not be enough to cha racterize the global geometry of the embedding space comprehensively. To address this problem, we perform linear interpolation on embeddings to adaptively manip ulate their hard levels and generate corresponding label-preserving synthetics f or recycled training, so that information buried in all samples can be fully exp loited and the metric is always challenged with proper difficulty. Our method ac hieves very competitive performance on the widely used CUB-200-2011, Cars196, an d Stanford Online Products datasets.

****

Auto-DeepLab: Hierarchical Neural Architecture Search for Semantic Image Segment ation

Chenxi Liu, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, Wei Hua, Alan L. Yuille, Li Fei-Fei; Proceedings of the IEEE/CVF Conference on Computer Visio n and Pattern Recognition (CVPR), 2019, pp. 82-92

Recently, Neural Architecture Search (NAS) has successfully identified neural ne twork architectures that exceed human designed ones on large-scale image classif ication. In this paper, we study NAS for semantic image segmentation. Existing w orks often focus on searching the repeatable cell structure, while hand-designin g the outer network structure that controls the spatial resolution changes. This choice simplifies the search space, but becomes increasingly problematic for de nse image prediction which exhibits a lot more network level architectural varia tions. Therefore, we propose to search the network level structure in addition t o the cell level structure, which forms a hierarchical architecture search space . We present a network level search space that includes many popular designs, an d develop a formulation that allows efficient gradient-based architecture search (3 P100 GPU days on Cityscapes images). We demonstrate the effectiveness of the proposed method on the challenging Cityscapes, PASCAL VOC 2012, and ADE20K data sets. Auto-DeepLab, our architecture searched specifically for semantic image se gmentation, attains state-of-the-art performance without any ImageNet pretrainin g.

****

Learning Loss for Active Learning

Donggeun Yoo, In So Kweon; Proceedings of the IEEE/CVF Conference on Computer V ision and Pattern Recognition (CVPR), 2019, pp. 93-102

The performance of deep neural networks improves with more annotated data. The p roblem is that the budget for annotation is limited. One solution to this is act ive learning, where a model asks human to annotate data that it perceived as unc ertain. A variety of recent methods have been proposed to apply active learning

to deep networks but most of them are either designed specific for their target tasks or computationally inefficient for large networks. In this paper, we propose a novel active learning method that is simple but task-agnostic, and works efficiently with the deep networks. We attach a small parametric module, named "loss prediction module," to a target network, and learn it to predict target losses of unlabeled inputs. Then, this module can suggest data that the target model is likely to produce a wrong prediction. This method is task-agnostic as networks are learned from a single loss regardless of target tasks. We rigorously validate our method through image classification, object detection, and human pose estimation, with the recent network architectures. The results demonstrate that our method consistently outperforms the previous methods over the tasks.
****

Striking the Right Balance With Uncertainty
Salman Khan, Munawar Hayat, Syed Waqas Zamir, Jianbing Shen, Ling Shao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 103-112
Learning unbiased models on imbalanced datasets is a significant challenge. Rare classes tend to get a concentrated representation in the classification space which hampers the generalization of learned boundaries to new test examples. In this paper, we demonstrate that the Bayesian uncertainty estimates directly correlate with the rarity of classes and the difficulty level of individual samples. Subsequently, we present a novel framework for uncertainty based class imbalance learning that follows two key insights: First, classification boundaries should be extended further away from a more uncertain (rare) class to avoid over-fitting and enhance its generalization. Second, each sample should be modeled as a multi-variate Gaussian distribution with a mean vector and a covariance matrix defined by the sample's uncertainty. The learned boundaries should respect not only the individual samples but also their distribution in the feature space. Our proposed approach efficiently utilizes sample and class uncertainty information to learn robust features and more generalizable classifiers. We systematically study the class imbalance problem and derive a novel loss formulation for max-margin learning based on Bayesian uncertainty measure. The proposed method shows significant performance improvements on six benchmark datasets for face verification, attribute prediction, digit/object classification and skin lesion detection.
****

AutoAugment: Learning Augmentation Strategies From Data
Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, Quoc V. Le; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 113-123
Data augmentation is an effective technique for improving the accuracy of modern image classifiers. However, current data augmentation implementations are manually designed. In this paper, we describe a simple procedure called AutoAugment to automatically search for improved data augmentation policies. In our implementation, we have designed a search space where a policy consists of many sub-policies, one of which is randomly chosen for each image in each mini-batch. A sub-policy consists of two operations, each operation being an image processing function such as translation, rotation, or shearing, and the probabilities and magnitudes with which the functions are applied. We use a search algorithm to find the best policy such that the neural network yields the highest validation accuracy on a target dataset. Our method achieves state-of-the-art accuracy on CIFAR-10, CIFAR-100, SVHN, and ImageNet (without additional data). On ImageNet, we attain a Top-1 accuracy of 83.5% which is 0.4% better than the previous record of 83.1%. On CIFAR-10, we achieve an error rate of 1.5%, which is 0.6% better than the previous state-of-the-art. Augmentation policies we find are transferable between datasets. The policy learned on ImageNet transfers well to achieve significant improvements on other datasets, such as Oxford Flowers, Caltech-101, Oxford-IIT Pets, FGVC Aircraft, and Stanford Cars.
****

SDRSAC: Semidefinite-Based Randomized Approach for Robust Point Cloud Registration Without Correspondences

Huu M. Le, Thanh-Toan Do, Tuan Hoang, Ngai-Man Cheung; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 124-133

This paper presents a novel randomized algorithm for robust point cloud registration without correspondences. Most existing registration approaches require a set of putative correspondences obtained by extracting invariant descriptors. However, such descriptors could become unreliable in noisy and contaminated settings. In these settings, methods that directly handle input point sets are preferable. Without correspondences, however, conventional randomized techniques require a very large number of samples in order to reach satisfactory solutions. In this paper, we propose a novel approach to address this problem. In particular, our work enables the use of randomized methods for point cloud registration without the need of putative correspondences. By considering point cloud alignment as a special instance of graph matching and employing an efficient semi-definite relaxation, we propose a novel sampling mechanism, in which the size of the sampled subsets can be larger-than-minimal. Our tight relaxation scheme enables fast rejection of the outliers in the sampled sets, resulting in high quality hypotheses. We conduct extensive experiments to demonstrate that our approach outperforms other state-of-the-art methods. Importantly, our proposed method serves as a generic framework which can be extended to problems with known correspondences.

****

BAD SLAM: Bundle Adjusted Direct RGB-D SLAM

Thomas Schops, Torsten Sattler, Marc Pollefeys; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 134-144

A key component of Simultaneous Localization and Mapping (SLAM) systems is the joint optimization of the estimated 3D map and camera trajectory. Bundle adjustment (BA) is the gold standard for this. Due to the large number of variables in dense RGB-D SLAM, previous work has focused on approximating BA. In contrast, in this paper we present a novel, fast direct BA formulation which we implement in a real-time dense RGB-D SLAM algorithm. In addition, we show that direct RGB-D SLAM systems are highly sensitive to rolling shutter, RGB and depth sensor synchronization, and calibration errors. In order to facilitate state-of-the-art research on direct RGB-D SLAM, we propose a novel, well-calibrated benchmark for this task that uses synchronized global shutter RGB and depth cameras. It includes a training set, a test set without public ground truth, and an online evaluation service. We observe that the ranking of methods changes on this dataset compared to existing ones, and our proposed algorithm outperforms all other evaluated SLAM methods. Our benchmark and our open source SLAM algorithm are available at: www.eth3d.net

****

Revealing Scenes by Inverting Structure From Motion Reconstructions

Francesco Pittaluga, Sanjeev J. Koppal, Sing Bing Kang, Sudipta N. Sinha; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 145-154

Many 3D vision systems localize cameras within a scene using 3D point clouds. Such point clouds are often obtained using structure from motion (SfM), after which the images are discarded to preserve privacy. In this paper, we show, for the first time, that such point clouds retain enough information to reveal scene appearance and compromise privacy. We present a privacy attack that reconstructs color images of the scene from the point cloud. Our method is based on a cascaded U-Net that takes as input, a 2D multichannel image of the points rendered from a specific viewpoint containing point depth and optionally color and SIFT descriptors and outputs a color image of the scene from that viewpoint. Unlike previous feature inversion methods, we deal with highly sparse and irregular 2D point distributions and inputs where many point attributes are missing, namely keypoint orientation and scale, the descriptor image source and the 3D point visibility. We evaluate our attack algorithm on public datasets and analyze the significance of the point cloud attributes. Finally, we show that novel views can also be generated thereby enabling compelling virtual tours of the underlying scene.

****

Strand-Accurate Multi-View Hair Capture
Giljoo Nam, Chenglei Wu, Min H. Kim, Yaser Sheikh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 155-164
Hair is one of the most challenging objects to reconstruct due to its micro-scale structure and a large number of repeated strands with heavy occlusions. In this paper, we present the first method to capture high-fidelity hair geometry with strand-level accuracy. Our method takes three stages to achieve this. In the first stage, a new multi-view stereo method with a slanted support line is proposed to solve the hair correspondences between different views. In detail, we contribute a novel cost function consisting of both photo-consistency term and geometric term that reconstructs each hair pixel as a 3D line. By merging all the depth maps, a point cloud, as well as local line directions for each point, is obtained. Thus, in the second stage, we feature a novel strand reconstruction method with the mean-shift to convert the noisy point data to a set of strands. Lastly, we grow the hair strands with multi-view geometric constraints to elongate the short strands and recover the missing strands, thus significantly increasing the reconstruction completeness. We evaluate our method on both synthetic data and real captured data, showing that our method can reconstruct hair strands with sub-millimeter accuracy.
****
DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation
Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, Steven Lovegrove; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 165-174
Computer graphics, 3D computer vision and robotics communities have produced multiple approaches to representing 3D geometry for rendering and reconstruction. These provide trade-offs across fidelity, efficiency and compression capabilities. In this work, we introduce DeepSDF, a learned continuous Signed Distance Function (SDF) representation of a class of shapes that enables high quality shape representation, interpolation and completion from partial and noisy 3D input data. DeepSDF, like its classical counterpart, represents a shape's surface by a continuous volumetric field: the magnitude of a point in the field represents the distance to the surface boundary and the sign indicates whether the region is inside (-) or outside (+) of the shape, hence our representation implicitly encodes a shape's boundary as the zero-level-set of the learned function while explicitly representing the classification of space as being part of the shapes interior or not. While classical SDF's both in analytical or discretized voxel form typically represent the surface of a single shape, DeepSDF can represent an entire class of shapes. Furthermore, we show state-of-the-art performance for learned 3D shape representation and completion while reducing the model size by an order of magnitude compared with previous work.
****
Pushing the Boundaries of View Extrapolation With Multiplane Images
Pratul P. Srinivasan, Richard Tucker, Jonathan T. Barron, Ravi Ramamoorthi, Ren Ng, Noah Snavely; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 175-184
We explore the problem of view synthesis from a narrow baseline pair of images, and focus on generating high-quality view extrapolations with plausible disocclusions. Our method builds upon prior work in predicting a multiplane image (MPI), which represents scene content as a set of RGBA planes within a reference view frustum and renders novel views by projecting this content into the target viewpoints. We present a theoretical analysis showing how the range of views that can be rendered from an MPI increases linearly with the MPI disparity sampling frequency, as well as a novel MPI prediction procedure that theoretically enables view extrapolations of up to 4 times the lateral viewpoint movement allowed by prior work. Our method ameliorates two specific issues that limit the range of views renderable by prior methods: 1) We expand the range of novel views that can be rendered without depth discretization artifacts by using a 3D convolutional network architecture along with a randomized-resolution training procedure to allow

our model to predict MPIs with increased disparity sampling frequency. 2) We reduce the repeated texture artifacts seen in disocclusions by enforcing a constraint that the appearance of hidden content at any depth must be drawn from visible content at or behind that depth.
****

GA-Net: Guided Aggregation Net for End-To-End Stereo Matching
Feihu Zhang, Victor Prisacariu, Ruigang Yang, Philip H.S. Torr; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 185-194
In the stereo matching task, matching cost aggregation is crucial in both traditional methods and deep neural network models in order to accurately estimate disparities. We propose two novel neural net layers, aimed at capturing local and the whole-image cost dependencies respectively. The first is a semi-global aggregation layer which is a differentiable approximation of the semi-global matching, the second is the local guided aggregation layer which follows a traditional cost filtering strategy to refine thin structures. These two layers can be used to replace the widely used 3D convolutional layer which is computationally costly and memory-consuming as it has cubic computational/memory complexity. In the experiments, we show that nets with a two-layer guided aggregation block easily outperform the state-of-the-art GC-Net which has nineteen 3D convolutional layers. We also train a deep guided aggregation network (GA-Net) which gets better accuracies than state-of-the-art methods on both Scene Flow dataset and KITTI benchmarks.
****

Real-Time Self-Adaptive Deep Stereo
Alessio Tonioni, Fabio Tosi, Matteo Poggi, Stefano Mattoccia, Luigi Di Stefano; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 195-204
Deep convolutional neural networks trained end-to-end are the state-of-the-art methods to regress dense disparity maps from stereo pairs. These models, however, suffer from a notable decrease in accuracy when exposed to scenarios significantly different from the training set (e.g., real vs synthetic images, etc.). We argue that it is extremely unlikely to gather enough samples to achieve effective training/tuning in any target domain, thus making this setup impractical for many applications. Instead, we propose to perform unsupervised and continuous online adaptation of a deep stereo network, which allows for preserving its accuracy in any environment. However, this strategy is extremely computationally demanding and thus prevents real-time inference. We address this issue introducing a new lightweight, yet effective, deep stereo architecture, Modularly ADaptive Network(MADNet), and developing a Modular ADaptation (MAD) algorithm, which independently trains sub-portions of the network. By deploying MADNet together with MAD we introduce the first real-time self-adaptive deep stereo system enabling competitive performance on heterogeneous datasets. Our code is publicly available at https://github.com/CVLAB-Unibo/Real-time-self-adaptive-deep-stereo.
****

LAF-Net: Locally Adaptive Fusion Networks for Stereo Confidence Estimation
Sunok Kim, Seungryong Kim, Dongbo Min, Kwanghoon Sohn; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 205-214
We present a novel method that estimates confidence map of an initial disparity by making full use of tri-modal input, including matching cost, disparity, and color image through deep networks. The proposed network, termed as Locally Adaptive Fusion Networks (LAF-Net), learns locally-varying attention and scale maps to fuse the tri-modal confidence features. The attention inference networks encode the importance of tri-modal confidence features and then concatenate them using the attention maps in an adaptive and dynamic fashion. This enables us to make an optimal fusion of the heterogeneous features, compared to a simple concatenation technique that is commonly used in conventional approaches. In addition, to encode the confidence features with locally-varying receptive fields, the scale inference networks learn the scale map and warp the fused confidence features t

hrough convolutional spatial transformer networks. Finally, the confidence map i s progressively estimated in the recursive refinement networks to enforce a spat ial context and local consistency. Experimental results show that this model out performs the state-of-the-art methods on various benchmarks.
****

NM-Net: Mining Reliable Neighbors for Robust Feature Correspondences
Chen Zhao, Zhiguo Cao, Chi Li, Xin Li, Jiaqi Yang; Proceedings of the IEEE/C VF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 215-2 24
Feature correspondence selection is pivotal to many feature-matching based tasks in computer vision. Searching spatially k-nearest neighbors is a common strateg y for extracting local information in many previous works. However, there is no guarantee that the spatially k-nearest neighbors of correspondences are consiste nt because the spatial distribution of false correspondences is often irregular. To address this issue, we present a compatibility-specific mining method to sea rch for consistent neighbors. Moreover, in order to extract and aggregate more r eliable features from neighbors, we propose a hierarchical network named NM-Net with a series of graph convolutions that is insensitive to the order of correspo ndences. Our experimental results have shown the proposed method achieves the st ate-of-the-art performance on four datasets with various inlier ratios and varyi ng numbers of feature consistencies.
****

Coordinate-Free Carlsson-Weinshall Duality and Relative Multi-View Geometry
Matthew Trager, Martial Hebert, Jean Ponce; Proceedings of the IEEE/CVF Confer ence on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 225-233
We present a coordinate-free description of Carlsson-Weinshall duality between s cene points and camera pinholes and use it to derive a new characterization of p rimal/dual multi-view geometry. In the case of three views, a particular set of reduced trilinearities provide a novel parameterization of camera geometry that, unlike existing ones, is subject only to very simple internal constraints. Thes e trilinearities lead to new "quasi-linear" algorithms for primal and dual struc ture from motion. We include some preliminary experiments with real and syntheti c data.
****

Deep Reinforcement Learning of Volume-Guided Progressive View Inpainting for 3D Point Scene Completion From a Single Depth Image
Xiaoguang Han, Zhaoxuan Zhang, Dong Du, Mingdai Yang, Jingming Yu, Pan Pan, Xin Yang, Ligang Liu, Zixiang Xiong, Shuguang Cui; Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 234-243
We present a deep reinforcement learning method of progressive view inpainting f or 3D point scene completion under volume guidance, achieving high-quality scene reconstruction from only a single depth image with severe occlusion. Our approa ch is end-to-end, consisting of three modules: 3D scene volume reconstruction, 2 D depth map inpainting, and multi-view selection for completion. Given a single depth image, our method first goes through the 3D volume branch to obtain a volu metric scene reconstruction as a guide to the next view inpainting step, which a ttempts to make up the missing information; the third step involves projecting t he volume under the same view of the input, concatenating them to complete the c urrent view depth, and integrating all depth into the point cloud. Since the occ luded areas are unavailable, we resort to a deep Q-Network to glance around and pick the next best view for large hole completion progressively until a scene is adequately reconstructed while guaranteeing validity. All steps are learned joi ntly to achieve robust and consistent results. We perform qualitative and quanti tative evaluations with extensive experiments on the SUNCG data, obtaining bette r results than the state of the art.
****

Video Action Transformer Network
Rohit Girdhar, Joao Carreira, Carl Doersch, Andrew Zisserman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019,

pp. 244-253
We introduce the Action Transformer model for recognizing and localizing human a ctions in video clips. We repurpose a Transformer-style architecture to aggregat e features from the spatiotemporal context around the person whose actions we ar e trying to classify. We show that by using high-resolution, person-specific, cl ass-agnostic queries, the model spontaneously learns to track individual people and to pick up on semantic context from the actions of others. Additionally its attention mechanism learns to emphasize hands and faces, which are often crucial to discriminate an action - all without explicit supervision other than boxes a nd class labels. We train and test our Action Transformer network on the Atomic Visual Actions (AVA) dataset, outperforming the state-of-the-art by a significan t margin using only raw RGB frames as input.
****

Timeception for Complex Action Recognition
Noureldien Hussein, Efstratios Gavves, Arnold W.M. Smeulders; Proceedings of t he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 254-263
This paper focuses on the temporal aspect for recognizing human activities in vi deos; an important visual cue that has long been undervalued. We revisit the con ventional definition of activity and restrict it to Complex Action: a set of one -actions with a weak temporal pattern that serves a specific purpose. Related wo rks use spatiotemporal 3D convolutions with fixed kernel size, too rigid to capt ure the varieties in temporal extents of complex actions, and too short for long -range temporal modeling. In contrast, we use multi-scale temporal convolutions, and we reduce the complexity of 3D convolutions. The outcome is Timeception con volution layers, which reasons about minute-long temporal patterns, a factor of 8 longer than best related works. As a result, Timeception achieves impressive a ccuracy in recognizing the human activities of Charades, Breakfast Actions and M ultiTHUMOS. Further, we demonstrate that Timeception learns long-range temporal dependencies and tolerate temporal extents of complex actions.
****

STEP: Spatio-Temporal Progressive Learning for Video Action Detection
Xitong Yang, Xiaodong Yang, Ming-Yu Liu, Fanyi Xiao, Larry S. Davis, Jan Ka utz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog nition (CVPR), 2019, pp. 264-272
In this paper, we propose Spatio-TEmporal Progressive (STEP) action detector--a progressive learning framework for spatio-temporal action detection in videos. S tarting from a handful of coarse-scale proposal cuboids, our approach progressiv ely refines the proposals towards actions over a few steps. In this way, high-qu ality proposals (i.e., adhere to action movements) can be gradually obtained at later steps by leveraging the regression outputs from previous steps. At each st ep, we adaptively extend the proposals in time to incorporate more related tempo ral context. Compared to the prior work that performs action detection in one ru n, our progressive learning framework is able to naturally handle the spatial di splacement within action tubes and therefore provides a more effective way for s patio-temporal modeling. We extensively evaluate our approach on UCF101 and AVA, and demonstrate superior detection results. Remarkably, we achieve mAP of 75.0% and 18.6% on the two datasets with 3 progressive steps and using respectively o nly 11 and 34 initial proposals.
****

Relational Action Forecasting
Chen Sun, Abhinav Shrivastava, Carl Vondrick, Rahul Sukthankar, Kevin Murphy , Cordelia Schmid; Proceedings of the IEEE/CVF Conference on Computer Vision an d Pattern Recognition (CVPR), 2019, pp. 273-283
This paper focuses on multi-person action forecasting in videos. More precisely, given a history of H previous frames, the goal is to detect actors and to predi ct their future actions for the next T frames. Our approach jointly models tempo ral and spatial interactions among different actors by constructing a recurrent graph, using actor proposals obtained with Faster R-CNN as nodes. Our method lea rns to select a subset of discriminative relations without requiring explicit su

pervision, thus enabling us to tackle challenging visual data. We refer to our model as Discriminative Relational Recurrent Network (DRRN). Evaluation of action prediction on AVA demonstrates the effectiveness of our proposed method compared to simpler baselines. Furthermore, we significantly improve performance on the task of early action classification on J-HMDB, from the previous SOTA of 48% to 60%.

****

Long-Term Feature Banks for Detailed Video Understanding

Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, Ross Girshick; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 284-293

To understand the world, we humans constantly need to relate the present to the past, and put events in context. In this paper, we enable existing video models to do the same. We propose a long-term feature bank--supportive information extracted over the entire span of a video--to augment state-of-the-art video models that otherwise would only view short clips of 2-5 seconds. Our experiments demonstrate that augmenting 3D convolutional networks with a long-term feature bank yields state-of-the-art results on three challenging video datasets: AVA, EPIC-Kitchens, and Charades. Code is available online.

****

Which Way Are You Going? Imitative Decision Learning for Path Forecasting in Dynamic Scenes

Yuke Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 294-303

Path forecasting is a pivotal step toward understanding dynamic scenes and an emerging topic in the computer vi- sion field. This task is challenging due to the multimodal nature of the future, namely, given a partial history, there is more than one plausible prediction. Yet, the state-of-the-art methods seem not fully responsive to this innate variabil- ity. Hence, how to better foresee the forthcoming trajectory in dynamic scenes has to be more thoroughly pursued. To this end, we propose a novel Imitative Decision Learning (IDL) approach. It delves deeper into the key that inher- ently characterizes the multimodality - the latent decision. The proposed IDL first infers the distribution of such latent decisions by learning from moving histories. A policy is then generated by taking the sampled latent decision into account to predict the future. Different plausible upcoming paths corresponds to each sampled latent decision. This ap- proach significantly differs from the mainstream literature that relies on a predefined latent variable to extrapolate di- verse predictions. In order to augment the understanding of the latent decision and resultant mutimodal future, we in- vestigate their connection through mutual information op- timization. Moreover, the proposed IDL integrates spatial and temporal dependencies into one single framework, in contrast to handling them with two-step settings. As a re- sult, our approach enables simultaneous anticipation of the paths of all pedestrians in the scene. We assess our pro- posal on the large-scale SAP, ETH and UCY datasets. The experiments show that IDL introduces considerable margin improvements with respect to recent leading studies.

****

What and How Well You Performed? A Multitask Learning Approach to Action Quality Assessment

Paritosh Parmar, Brendan Tran Morris; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 304-313

Can performance on the task of action quality assessment (AQA) be improved by exploiting a description of the action and its quality? Current AQA and skills assessment approaches propose to learn features that serve only one task - estimating the final score. In this paper, we propose to learn spatio-temporal features that explain three related tasks - fine-grained action recognition, commentary generation, and estimating the AQA score. A new multitask-AQA dataset, the largest to date, comprising of 1412 diving samples was collected to evaluate our approach (http://rtis.oit.unlv.edu/datasets.html). We show that our MTL approach outperforms STL approach using two different kinds of architectures: C3D-AVG and MSC

ADC. The C3D-AVG-MTL approach achieves the new state-of-the-art performance with a rank correlation of 90.44%. Detailed experiments were performed to show that MTL offers better generalization than STL, and representations from action recognition models are not sufficient for the AQA task and instead should be learned.
****

MHP-VOS: Multiple Hypotheses Propagation for Video Object Segmentation
Shuangjie Xu, Daizong Liu, Linchao Bao, Wei Liu, Pan Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 314-323
We address the problem of semi-supervised video object segmentation (VOS), where the masks of objects of interests are given in the first frame of an input video. To deal with challenging cases where objects are occluded or missing, previous work relies on greedy data association strategies that make decisions for each frame individually. In this paper, we propose a novel approach to defer the decision making for a target object in each frame, until a global view can be established with the entire video being taken into consideration. Our approach is in the same spirit as Multiple Hypotheses Tracking (MHT) methods, making several critical adaptations for the VOS problem. We employ the bounding box (bbox) hypothesis for tracking tree formation, and the multiple hypotheses are spawned by propagating the preceding bbox into the detected bbox proposals within a gated region starting from the initial object mask in the first frame. The gated region is determined by a gating scheme which takes into account a more comprehensive motion model rather than the simple Kalman filtering model in traditional MHT. To further design more customized algorithms tailored for VOS, we develop a novel mask propagation score instead of the appearance similarity score that could be brittle due to large deformations. The mask propagation score, together with the motion score, determines the affinity between the hypotheses during tree pruning. Finally, a novel mask merging strategy is employed to handle mask conflicts between objects. Extensive experiments on challenging datasets demonstrate the effectiveness of the proposed method, especially in the case of object missing.
****

2.5D Visual Sound
Ruohan Gao, Kristen Grauman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 324-333
Binaural audio provides a listener with 3D sound sensation, allowing a rich perceptual experience of the scene. However, binaural recordings are scarcely available and require nontrivial expertise and equipment to obtain. We propose to convert common monaural audio into binaural audio by leveraging video. The key idea is that visual frames reveal significant spatial cues that, while explicitly lacking in the accompanying single-channel audio, are strongly linked to it. Our multi-modal approach recovers this link from unlabeled video. We devise a deep convolutional neural network that learns to decode the monaural (single-channel) soundtrack into its binaural counterpart by injecting visual information about object and scene configurations. We call the resulting output 2.5D visual sound---the visual stream helps "lift" the flat single channel audio into spatialized sound. In addition to sound generation, we show the self-supervised representation learned by our network benefits audio-visual source separation. Our video results: http://vision.cs.utexas.edu/projects/2.5D_visual_sound/
****

Language-Driven Temporal Activity Localization: A Semantic Matching Reinforcement Learning Model
Weining Wang, Yan Huang, Liang Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 334-343
Current studies on action detection in untrimmed videos are mostly designed for action classes, where an action is described at word level such as jumping, tumbling, swing, etc. This paper focuses on a rarely investigated problem of localizing an activity via a sentence query which would be more challenging and practical. Considering that current methods are generally time-consuming due to the dense frame-processing manner, we propose a recurrent neural network based reinforcement learning model which selectively observes a sequence of frames and associa

tes the given sentence with video content in a matching-based manner. However, directly matching sentences with video content performs poorly due to the large visual-semantic discrepancy. Thus, we extend the method to a semantic matching reinforcement learning (SM-RL) model by extracting semantic concepts of videos and then fusing them with global context features. Extensive experiments on three benchmark datasets, TACoS, Charades-STA and DiDeMo, show that our method achieves the state-of-the-art performance with a high detection speed, demonstrating both effectiveness and efficiency of our method.
****

Gaussian Temporal Awareness Networks for Action Localization
Fuchen Long, Ting Yao, Zhaofan Qiu, Xinmei Tian, Jiebo Luo, Tao Mei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 344-353
Temporally localizing actions in a video is a fundamental challenge in video understanding. Most existing approaches have often drawn inspiration from image object detection and extended the advances, e.g., SSD and Faster R-CNN, to produce temporal locations of an action in a 1D sequence. Nevertheless, the results can suffer from robustness problem due to the design of predetermined temporal scales, which overlooks the temporal structure of an action and limits the utility on detecting actions with complex variations. In this paper, we propose to address the problem by introducing Gaussian kernels to dynamically optimize temporal scale of each action proposal. Specifically, we present Gaussian Temporal Awareness Networks (GTAN) --- a new architecture that novelly integrates the exploitation of temporal structure into an one-stage action localization framework. Technically, GTAN models the temporal structure through learning a set of Gaussian kernels, each for a cell in the feature maps. Each Gaussian kernel corresponds to a particular interval of an action proposal and a mixture of Gaussian kernels could further characterize action proposals with various length. Moreover, the values in each Gaussian curve reflect the contextual contributions to the localization of an action proposal. Extensive experiments are conducted on both THUMOS14 and ActivityNet v1.3 datasets, and superior results are reported when comparing to state-of-the-art approaches. More remarkably, GTAN achieves 1.9% and 1.1% improvements in mAP on testing set of the two datasets.
****

Efficient Video Classification Using Fewer Frames
Shweta Bhardwaj, Mukundhan Srinivasan, Mitesh M. Khapra; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 354-363
Recently, there has been a lot of interest in building compact models for video classification which have a small memory footprint (<1 GB). While these models are compact, they typically operate by repeated application of a small weight matrix to all the frames in a video. For example, recurrent neural network based methods compute a hidden state for every frame of the video using a recurrent weight matrix. Similarly, cluster-and-aggregate based methods such as NetVLAD have a learnable clustering matrix which is used to assign soft-clusters to every frame in the video. Since these models look at every frame in the video, the number of floating point operations (FLOPs) is still large even though the memory footprint is small. In this work, we focus on building compute-efficient video classification models which process fewer frames and hence have less number of FLOPs. Similar to memory efficient models, we use the idea of distillation albeit in a different setting. Specifically, in our case, a compute-heavy teacher which looks at all the frames in the video is used to train a compute-efficient student which looks at only a small fraction of frames in the video. This is in contrast to a typical memory efficient Teacher-Student setting, wherein both the teacher and the student look at all the frames in the video but the student has fewer parameters. Our work thus complements the research on memory efficient video classification. We do an extensive evaluation with three types of models for video classification, viz., (i) recurrent models (ii) cluster-and-aggregate models and (iii) memory-efficient cluster-and-aggregate models and show that in each of these cases, a see-it-all teacher can be used to train a compute efficient see-very-l

ittle student. Overall, we show that the proposed student network can reduce th e inference time by 30% and the number of FLOPs by approximately 90% with a negl igent drop in the performance.
****

Parsing R-CNN for Instance-Level Human Analysis
Lu Yang, Qing Song, Zhihui Wang, Ming Jiang; Proceedings of the IEEE/CVF Conf erence on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 364-373
Instance-level human analysis is common in real-life scenarios and has multiple manifestations, such as human part segmentation, dense pose estimation, human-ob ject interactions, etc. Models need to distinguish different human instances in the image panel and learn rich features to represent the details of each instanc e. In this paper, we present an end-to-end pipeline for solving the instance-lev el human analysis, named Parsing R-CNN. It processes a set of human instances si multaneously through comprehensive considering the characteristics of region-bas ed approach and the appearance of a human, thus allowing representing the detail s of instances. Parsing R-CNN is very flexible and efficient, which is appli cable to many issues in human instance analysis. Our approach outperforms all st ate-of-the-art methods on CIHP (Crowd Instance-level Human Parsing), MHP v2.0 (M ulti-Human Parsing) and DensePose-COCO datasets. Based on the proposed Parsing R -CNN, we reach the 1st place in the COCO 2018 Challenge DensePose Estimation tas k. Code and models are publicly available.
****

Large Scale Incremental Learning
Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, Yun Fu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern R ecognition (CVPR), 2019, pp. 374-382
Modern machine learning suffers from catastrophic forgetting when learning new c lasses incrementally. The performance dramatically degrades due to the missing d ata of old classes. Incremental learning methods have been proposed to retain th e knowledge acquired from the old classes, by using knowledge distilling and kee ping a few exemplars from the old classes. However, these methods struggle to sc ale up to a large number of classes. We believe this is because of the combinati on of two factors: (a) the data imbalance between the old and new classes, and (b) the increasing number of visually similar classes. Distinguishing betwee n an increasing number of visually similar classes is particularly challenging, when the training data is unbalanced. We propose a simple and effective method to address this data imbalance issue. We found that the last fully connected la yer has a strong bias towards the new classes, and this bias can be corrected by a linear model. With two bias parameters, our method performs remarkably wel l on two large datasets: ImageNet (1000 classes) and MS-Celeb-1M (10000 classes) , outperforming the state-of-the-art algorithms by 11.1% and 13.2% respectively.
****

TopNet: Structural Point Cloud Decoder
Lyne P. Tchapmi, Vineet Kosaraju, Hamid Rezatofighi, Ian Reid, Silvio Savare se; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn ition (CVPR), 2019, pp. 383-392
3D point cloud generation is of great use for 3D scene modeling and understandin g. Real-world 3D object point clouds can be properly described by a coll ection of low-level and high-level structures such as surfaces, geometric prim itives, semantic parts,etc. In fact, there exist many different representations of a 3D object point cloud as a set of point groups. Existing frameworks for po int cloud genera-ion either do not consider structure in their proposed solution s, or assume and enforce a specific structure/topology,e.g. a collection of manifolds or surfaces, for the generated point cloud of a 3D object. In this work, we pro-pose a novel decoder that generates a structured poin t cloud without assuming any specific structure or topology on the underlying po int set. Our decoder is softly constrained to generate a point cloud following a hierarchical rooted tree structure. We show that given enough capacity and al lowing for redundancies, the proposed decoder is very flexible and able to learn any arbitrary grouping of points including any topology on the point set. We e

valuate our decoder on the task of point cloud generation for 3D point cloud sha pe completion. Combined with encoders from existing frameworks, we show that ou r proposed decoder significantly outperforms state-of-the-art 3D point cloud com pletion methods on the Shapenet dataset
****

Perceive Where to Focus: Learning Visibility-Aware Part-Level Features for Parti al Person Re-Identification
Yifan Sun, Qin Xu, Yali Li, Chi Zhang, Yikang Li, Shengjin Wang, Jian Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti on (CVPR), 2019, pp. 393-402
This paper considers a realistic problem in person re-identification (re-ID) tas k, i.e., partial re-ID. Under partial re-ID scenario, the images may contain a p artial observation of a pedestrian. If we directly compare a partial pedestrian image with a holistic one, the extreme spatial misalignment significantly compro mises the discriminative ability of the learned representation. We propose a Vis ibility-aware Part Model (VPM) for partial re-ID, which learns to perceive the v isibility of regions through self-supervision. The visibility awareness allows V PM to extract region-level features and compare two images with focus on their s hared regions (which are visible on both images). VPM gains two-fold benefit tow ard higher accuracy for partial re-ID. On the one hand, compared with learning a global feature, VPM learns region-level features and thus benefits from fine-gr ained information. On the other hand, with visibility awareness, VPM is capable to estimate the shared regions between two images and thus suppresses the spatia l misalignment. Experimental results confirm that our method significantly impro ves the learned feature representation and the achieved accuracy is on par with the state of the art.
****

Meta-Transfer Learning for Few-Shot Learning
Qianru Sun, Yaoyao Liu, Tat-Seng Chua, Bernt Schiele; Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 403 -412
Meta-learning has been proposed as a framework to address the challenging few-sh ot learning setting. The key idea is to leverage a large number of similar few-s hot tasks in order to learn how to adapt a base-learner to a new task for which only a few labeled samples are available. As deep neural networks (DNNs) tend to overfit using a few samples only, meta-learning typically uses shallow neural n etworks (SNNs), thus limiting its effectiveness. In this paper we propose a nove l few-shot learning method called meta-transfer learning (MTL) which learns to a dapt a deep NN for few shot learning tasks. Specifically, "meta" refers to train ing multiple tasks, and "transfer" is achieved by learning scaling and shifting functions of DNN weights for each task. In addition, we introduce the hard task (HT) meta-batch scheme as an effective learning curriculum for MTL. We conduct e xperiments using (5-class, 1-shot) and (5-class, 5-shot) recognition tasks on tw o challenging few-shot learning benchmarks: miniImageNet and Fewshot-CIFAR100. E xtensive comparisons to related works validate that our meta-transfer learning a pproach trained with the proposed HT meta-batch scheme achieves top performance. An ablation study also shows that both components contribute to fast convergenc e and high accuracy.
****

Structured Binary Neural Networks for Accurate Image Classification and Semantic Segmentation
Bohan Zhuang, Chunhua Shen, Mingkui Tan, Lingqiao Liu, Ian Reid; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2 019, pp. 413-422
In this paper, we propose to train convolutional neural networks (CNNs) with bot h binarized weights and activations, leading to quantized models specifically fo r mobile devices with limited power capacity and computation resources. By assum ing the same architecture to full-precision networks, previous works on quantizi ng CNNs seek to preserve the floating-point information using a set of discrete values, which we call value approximation. However, we take a novel "structure a

pproximation" view for quantization--- it is very likely that a different archit ecture may be better for best performance. In particular, we propose a "network decomposition" strategy, named Group-Net, in which we divide the network into gr oups. In this way, each full-precision group can be effectively reconstructed by aggregating a set of homogeneous binary branches. In addition, we learn effect connections among groups to improve the representational capability. Moreover, the proposed Group-Net shows strong generalization to other tasks. For instance, we extend Group-Net for highly accurate semantic segmentation by embedding rich context into the binary structure. Experiments on both classification and seman tic segmentation tasks demonstrate the superior performance of the proposed meth ods over various popular architectures. In particular, we outperform the previou s best binary neural networks in terms of accuracy and huge computation saving.
****

Deep RNN Framework for Visual Sequential Applications

Bo Pang, Kaiwen Zha, Hanwen Cao, Chen Shi, Cewu Lu; Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 423-432

Extracting temporal and representation features efficiently plays a pivotal role in understanding visual sequence information. To deal with this, we propose a n ew recurrent neural framework that can be stacked deep effectively. There are ma inly two novel designs in our deep RNN framework: one is a new RNN module called Context Bridge Module (CBM) which splits the information flowing along the sequ ence (temporal direction) and along depth (spatial representation direction), ma king it easier to train when building deep by balancing these two directions; th e other is the Overlap Coherence Training Scheme that reduces the training compl exity for long visual sequential tasks on account of the limitation of computing resources. We provide empirical evidence to show that our deep RNN framework is easy to optimize and can gain accuracy from the increased depth on several vi sual sequence problems. On these tasks, we evaluate our deep RNN framework with 15 layers, 7x than conventional RNN networks, but it is still easy to train. Our deep framework achieves more than 11% relative improvements over shallow RNN mo dels on Kinetics, UCF-101, and HMDB-51 for video classification. For auxiliary a nnotation, after replacing the shallow RNN part of Polygon-RNN with our 15-layer deep CBM, the performance improves by 14.7%. For video future prediction, our d eep RNN improves the state-of-the-art shallow model's performance by 2.4% on PSN R and SSIM.
****

Graph-Based Global Reasoning Networks

Yunpeng Chen, Marcus Rohrbach, Zhicheng Yan, Yan Shuicheng, Jiashi Feng, Ya nnis Kalantidis; Proceedings of the IEEE/CVF Conference on Computer Vision and P attern Recognition (CVPR), 2019, pp. 433-442

Globally modeling and reasoning over relations between regions can be beneficial for many computer vision tasks on both images and videos. Convolutional Neural Networks (CNNs) excel at modeling local relations by convolution operations, but they are typically inefficient at capturing global relations between distant re gions and require stacking multiple convolution layers. In this work, we propose a new approach for reasoning globally in which a set of features are globally a ggregated over the coordinate space and then projected to an interaction space w here relational reasoning can be efficiently computed. After reasoning, relation -aware features are distributed back to the original coordinate space for down-s tream tasks. We further present a highly efficient instantiation of the proposed approach and introduce the Global Reasoning unit (GloRe unit) that implements t he coordinate-interaction space mapping by weighted global pooling and weighted broadcasting, and the relation reasoning via graph convolution on a small graph in interaction space. The proposed GloRe unit is lightweight, end-to-end trainab le and can be easily plugged into existing CNNs for a wide range of tasks. Exten sive experiments show our GloRe unit can consistently boost the performance of s tate-of-the-art backbone architectures, including ResNet, ResNeXt, SE-Net and DP N, for both 2D and 3D CNNs, on image classification, semantic segmentation and v ideo action recognition task.

****
SSN: Learning Sparse Switchable Normalization via SparsestMax
Wenqi Shao, Tianjian Meng, Jingyu Li, Ruimao Zhang, Yudian Li, Xiaogang Wang, Ping Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 443-451
Normalization methods improve both optimization and generalization of ConvNets. To further boost performance, the recently-proposed switchable normalization (SN) provides a new perspective for deep learning: it learns to select different normalizers for different convolution layers of a ConvNet. However, SN uses softmax function to learn importance ratios to combine normalizers, leading to redundant computations compared to a single normalizer. This work addresses this issue by presenting Sparse Switchable Normalization (SSN) where the importance ratios are constrained to be sparse. Unlike l_1 and l_0 constraints that impose difficulties in optimization, we turn this constrained optimization problem into feed-forward computation by proposing SparsestMax, which is a sparse version of softmax. SSN has several appealing properties. (1) It inherits all benefits from SN such as applicability in various tasks and robustness to a wide range of batch sizes. (2) It is guaranteed to select only one normalizer for each normalization layer, avoiding redundant computations. (3) SSN can be transferred to various tasks in an end-to-end manner. Extensive experiments show that SSN outperforms its counterparts on various challenging benchmarks such as ImageNet, Cityscapes, ADE20K, and Kinetics. Code is available at https://github.com/switchablenorms/Sparse_SwitchNorm.
****
Spherical Fractal Convolutional Neural Networks for Point Cloud Recognition
Yongming Rao, Jiwen Lu, Jie Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 452-460
We present a generic, flexible and 3D rotation invariant framework based on spherical symmetry for point cloud recognition. By introducing regular icosahedral lattice and its fractals to approximate and discretize sphere, convolution can be easily implemented to process 3D points. Based on the fractal structure, a hierarchical feature learning framework together with an adaptive sphere projection module is proposed to learn deep feature in an end-to-end manner. Our framework not only inherits the strong representation power and generalization capability from convolutional neural networks for image recognition, but also extends CNN to learn robust feature resistant to rotations and perturbations. The proposed model is effective yet robust. Comprehensive experimental study demonstrates that our approach can achieve competitive performance compared to state-of-the-art techniques on both 3D object classification and part segmentation tasks, meanwhile, outperform other rotation invariant models on rotated 3D object classification and retrieval tasks by a large margin.
****
Learning to Generate Synthetic Data via Compositing
Shashank Tripathi, Siddhartha Chandra, Amit Agrawal, Ambrish Tyagi, James M. Rehg, Visesh Chari; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 461-470
We present a task-specific approach to synthetic data generation. Our framework employs a trainable synthesizer network that is optimized to produce meaningful training samples by assessing the strengths and weaknesses of a 'target' classifier. The synthesizer and target networks are trained in an adversarial manner wherein each network is updated with a goal to outdo the other. Additionally, we ensure the synthesizer generates realistic data by pairing it with a discriminator trained on real-world images. Further, to make the target classifier invariant to blending artefacts, we introduce these artefacts to background regions of the training images so the target does not over-fit to them. We demonstrate the efficacy of our approach by applying it to different target networks including a classification network on AffNIST [46], and two object detection networks (SSD, Faster-RCNN) on different datasets. On the AffNIST benchmark, our approach is able to surpass the baseline results with just half the training examples. On the VOC person detection benchmark, we show improvements of up to 2.7% as a result o

f our data augmentation. Similarly on the GMU detection benchmark, we report a p erformance boost of 3.5% in mAP over the baseline method, outperforming the prev ious state of the art approaches by as much as 7.5% in individual categories.
****

Divide and Conquer the Embedding Space for Metric Learning
Artsiom Sanakoyeu, Vadim Tschernezki, Uta Buchler, Bjorn Ommer; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 201 9, pp. 471-480
Learning the embedding space, where semantically similar objects are located clo se together and dissimilar objects far apart, is a cornerstone of many computer vision applications. Existing approaches usually learn a single metric in the embedding space for all available data points, which may have a very complex non -uniform distribution with different notions of similarity between objects, e.g. appearance, shape, color or semantic meaning. Approaches for learning a single distance metric often struggle to encode all different types of relationships an d do not generalize well. In this work, we propose a novel easy-to-implement d ivide and conquer approach for deep metric learning, which significantly improv es the state-of-the-art performance of metric learning. Our approach utilizes th e embedding space more efficiently by jointly splitting the embedding space and data into K smaller sub-problems. It divides both, the data and the embedding s pace into K subsets and learns K separate distance metrics in the non-overlappin g subspaces of the embedding space, defined by groups of neurons in the embeddin g layer of the neural network. The proposed approach increases the convergence s peed and improves generalization since the complexity of each sub-problem is red uced compared to the original one. We show that our approach outperforms the sta te-of-the-art by a large margin in retrieval, clustering and re-identification t asks on CUB200-2011, CARS196, Stanford Online Products, In-shop Clothes and PKU VehicleID datasets. Source code: https://bit.ly/dcesml.
****

Latent Space Autoregression for Novelty Detection
Davide Abati, Angelo Porrello, Simone Calderara, Rita Cucchiara; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20 19, pp. 481-490
Novelty detection is commonly referred as the discrimination of observations tha t do not conform to a learned model of regularity. Despite its importance in dif ferent application settings, designing a novelty detector is utterly complex due to the unpredictable nature of novelties and its inaccessibility during the tra ining procedure, factors which expose the unsupervised nature of the problem. In our proposal, we design a general unsupervised framework where we equip a deep autoencoder with a parametric density estimator that learns the probability dist ribution underlying the latent representations with an autoregressive procedure. We show that a maximum likelihood objective, optimized in conjunction with the reconstruction of normal samples, effectively acts as a regularizer for the tas k at hand, by minimizing the differential entropy of the distribution spanned by latent vectors. In addition to providing a very general formulation, extensive experiments of our model on publicly available datasets deliver on-par or superi or performances if compared to state-of-the-art methods in one-class and in vide o anomaly detection settings. Differently from our competitors, we remark that o ur proposal does not make any assumption about the nature of the novelties, maki ng our work easily applicable to disparate contexts.
****

Attending to Discriminative Certainty for Domain Adaptation
Vinod Kumar Kurmi, Shanu Kumar, Vinay P. Namboodiri; Proceedings of the IEEE/C VF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 491-5 00
In this paper, we aim to solve for unsupervised domain adaptation of classifiers where we have access to label information for the source domain while these are not available for a target domain. While various methods have been proposed for solving these including adversarial discriminator based methods, most approache s have focused on the entire image based domain adaptation. In an image, there w

ould be regions that can be adapted better, for instance, the foreground object may be similar in nature. To obtain such regions, we propose methods that consider the probabilistic certainty estimate of various regions and specific focus on these during classification for adaptation. We observe that just by incorporating the probabilistic certainty of the discriminator while training the classifier, we are able to obtain state of the art results on various datasets as compared against all the recent methods. We provide a thorough empirical analysis of the method by providing ablation analysis, statistical significance test, and visualization of the attention maps and t-SNE embeddings. These evaluations convincingly demonstrate the effectiveness of the proposed approach.
****

Feature Denoising for Improving Adversarial Robustness
Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L. Yuille, Kaiming He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 501-509
Adversarial attacks to image classification systems present challenges to convolutional networks and opportunities for understanding them. This study suggests that adversarial perturbations on images lead to noise in the features constructed by these networks. Motivated by this observation, we develop new network architectures that increase adversarial robustness by performing feature denoising. Specifically, our networks contain blocks that denoise the features using non-local means or other filters; the entire networks are trained end-to-end. When combined with adversarial training, our feature denoising networks substantially improve the state-of-the-art in adversarial robustness in both white-box and black-box attack settings. On ImageNet, under 10-iteration PGD white-box attacks where prior art has 27.9% accuracy, our method achieves 55.7%; even under extreme 2000-iteration PGD white-box attacks, our method secures 42.6% accuracy. Our method was ranked first in Competition on Adversarial Attacks and Defenses (CAAD) 2018 --- it achieved 50.6% classification accuracy on a secret, ImageNet-like test dataset against 48 unknown attackers, surpassing the runner-up approach by 10%. Code is available at https://github.com/facebookresearch/ImageNet-Adversarial-Training.
****

Selective Kernel Networks
Xiang Li, Wenhai Wang, Xiaolin Hu, Jian Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 510-519
In standard Convolutional Neural Networks (CNNs), the receptive fields of artificial neurons in each layer are designed to share the same size. It is well-known in the neuroscience community that the receptive field size of visual cortical neurons are modulated by the stimulus, which has been rarely considered in constructing CNNs. We propose a dynamic selection mechanism in CNNs that allows each neuron to adaptively adjust its receptive field size based on multiple scales of input information. A building block called Selective Kernel (SK) unit is designed, in which multiple branches with different kernel sizes are fused using softmax attention that is guided by the information in these branches. Different attentions on these branches yield different sizes of the effective receptive fields of neurons in the fusion layer. Multiple SK units are stacked to a deep network termed Selective Kernel Networks (SKNets). On the ImageNet and CIFAR benchmarks, we empirically show that SKNet outperforms the existing state-of-the-art architectures with lower model complexity. Detailed analyses show that the neurons in SKNet can capture target objects with different scales, which verifies the capability of neurons for adaptively adjusting their receptive field sizes according to the input. The code and models are available at https://github.com/implus/SKNet.
****

On Implicit Filter Level Sparsity in Convolutional Neural Networks
Dushyant Mehta, Kwang In Kim, Christian Theobalt; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 520-528
We investigate filter level sparsity that emerges in convolutional neural networks (CNNs) which employ Batch Normalization and ReLU activation, and are trained

with adaptive gradient descent techniques and L2 regularization or weight decay. We conduct an extensive experimental study casting our initial findings into hypotheses and conclusions about the mechanisms underlying the emergent filter level sparsity. This study allows new insight into the performance gap obeserved between adapative and non-adaptive gradient descent methods in practice. Further, analysis of the effect of training strategies and hyperparameters on the sparsity leads to practical suggestions in designing CNN training strategies enabling us to explore the tradeoffs between feature selectivity, network capacity, and generalization performance. Lastly, we show that the implicit sparsity can be harnessed for neural network speedup at par or better than explicit sparsification / pruning approaches, with no modifications to the typical training pipeline required.

****

FlowNet3D: Learning Scene Flow in 3D Point Clouds

Xingyu Liu, Charles R. Qi, Leonidas J. Guibas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 529-537

Many applications in robotics and human-computer interaction can benefit from understanding 3D motion of points in a dynamic environment, widely noted as scene flow. While most previous methods focus on stereo and RGB-D images as input, few try to estimate scene flow directly from point clouds. In this work, we propose a novel deep neural network named FlowNet3D that learns scene flow from point clouds in an end-to-end fashion. Our network simultaneously learns deep hierarchical features of point clouds and flow embeddings that represent point motions, supported by two newly proposed learning layers for point sets. We evaluate the network on both challenging synthetic data from FlyingThings3D and real Lidar scans from KITTI. Trained on synthetic data only, our network successfully generalizes to real scans, outperforming various baselines and showing competitive results to the prior art. We also demonstrate two applications of our scene flow output (scan registration and motion segmentation) to show its potential wide use cases.

****

Scene Memory Transformer for Embodied Agents in Long-Horizon Tasks

Kuan Fang, Alexander Toshev, Li Fei-Fei, Silvio Savarese; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 538-547

Many robotic applications require the agent to perform long-horizon tasks in partially observable environments. In such applications, decision making at any step can depend on observations received far in the past. Hence, being able to properly memorize and utilize the long-term history is crucial. In this work, we propose a novel memory-based policy, named Scene Memory Transformer (SMT). The proposed policy embeds and adds each observation to a memory and uses the attention mechanism to exploit spatio-temporal dependencies. This model is generic and can be efficiently trained with reinforcement learning over long episodes. On a range of visual navigation tasks, SMT demonstrates superior performance to existing reactive and memory-based policies by a margin.

****

Co-Occurrent Features in Semantic Segmentation

Hang Zhang, Han Zhang, Chenguang Wang, Junyuan Xie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 548-557

Recent work has achieved great success in utilizing global contextual information for semantic segmentation, including increasing the receptive field and aggregating pyramid feature representations. In this paper, we go beyond global context and explore the fine-grained representation using co-occurrent features by introducing Co-occurrent Feature Model, which predicts the distribution of co-occurrent features for a given target. To leverage the semantic context in the co-occurrent features, we build an Aggregated Co-occurrent Feature (ACF) Module by aggregating the probability of the co-occurrent feature with the co-occurrent context. ACF Module learns a fine-grained spatial invariant representation to capture co-occurrent context information across the scene. Our approach significantly

improves the segmentation results using FCN and achieves superior performance 54 .0% mIoU on Pascal Context, 87.2% mIoU on Pascal VOC 2012 and 44.89% mIoU on ADE 20K datasets. The source code and complete system will be publicly available upo n publication.
****

Bag of Tricks for Image Classification with Convolutional Neural Networks
Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, Mu Li; Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 558-567
Much of the recent progress made in image classification research can be credite d to training procedure refinements, such as changes in data augmentations and optimization methods. In the literature, however, most refinements are either br iefly mentioned as implementation details or only visible in source code. In thi s paper, we will examine a collection of such refinements and empirically evalua te their impact on the final model accuracy through ablation study. We will show that, by combining these refinements together, we are able to improve various C NN models significantly. For example, we raise ResNet-50's top-1 validation accu racy from 75.3% to 79.29% on ImageNet. We will also demonstrate that improvement on image classification accuracy leads to better transfer learning performance in other application domains such as object detection and semantic segmentation.
****

Learning Channel-Wise Interactions for Binary Convolutional Neural Networks
Ziwei Wang, Jiwen Lu, Chenxin Tao, Jie Zhou, Qi Tian; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 56 8-577
In this paper, we propose a channel-wise interaction based binary convolutional neural network learning method (CI-BCNN) for efficient inference. Conventional m ethods apply xnor and bitcount operations in binary convolution with notable qua ntization error, which usually obtains inconsistent signs in binary feature maps compared with their full-precision counterpart and leads to significant informa tion loss. In contrast, our CI-BCNN mines the channel-wise interactions, through which prior knowledge is provided to alleviate inconsistency of signs in binary feature maps and preserves the information of input samples during inference. S pecifically, we mine the channel-wise interactions by a reinforcement learning m odel, and impose channel-wise priors on the intermediate feature maps through th e interacted bitcount function. Extensive experiments on the CIFAR-10 and ImageN et datasets show that our method outperforms the state-of-the-art binary convolu tional neural networks with less computational and storage cost.
****

Knowledge Adaptation for Efficient Semantic Segmentation
Tong He, Chunhua Shen, Zhi Tian, Dong Gong, Changming Sun, Youliang Yan; Pr oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 578-587
Both accuracy and efficiency are of significant importance to the task of semant ic segmentation. Existing deep FCNs suffer from heavy computations due to a seri es of high-resolution feature maps for preserving the detailed knowledge in dens e estimation. Although reducing the feature map resolution (i.e., applying a lar ge overall stride) via subsampling operations (e.g., polling and convolution str iding) can instantly increase the efficiency, it dramatically decreases the esti mation accuracy. To tackle this dilemma, we propose a knowledge distillation met hod tailored for semantic segmentation to improve the performance of the compact FCNs with large overall stride. To handle the inconsistency between the feature s of the student and teacher network, we optimize the feature similarity in a tr ansferred latent domain formulated by utilizing a pre-trained autoencoder. Moreo ver, an affinity distillation module is proposed to capture the long-range depen dency by calculating the non local interactions across the whole image. To valid ate the effectiveness of our proposed method, extensive experiments have been co nducted on three popular benchmarks: Pascal VOC, Cityscapes and Pascal Context. Built upon a highly competitive baseline, our proposed method can improve the pe rformance of a student network by 2.5% (mIOU boosts from 70.2 to 72.7 on the cit

yscapes test set) and can train a better compact model with only 8% float operat
ions (FLOPS) of a model that achieves comparable performances.
****

Parametric Noise Injection: Trainable Randomness to Improve Deep Neural Network
Robustness Against Adversarial Attack
Zhezhi He, Adnan Siraj Rakin, Deliang Fan; Proceedings of the IEEE/CVF Confere
nce on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 588-597
Recent developments in the field of Deep Learning have exposed the underlying vu
lnerability of Deep Neural Network (DNN) against adversarial examples. In image
classification, an adversarial example is a carefully modified image that is vis
ually imperceptible to the original image but can cause DNN model to misclassify
 it. Training the network with Gaussian noise is an effective technique to perfo
rm model regularization, thus improving model robustness against input variation
. Inspired by this classical method, we explore to utilize the regularization ch
aracteristic of noise injection to improve DNN's robustness against adversarial
attack. In this work, we propose Parametric-Noise-Injection (PNI) which involves
 trainable Gaussian noise injection at each layer on either activation or weight
s through solving the Min-Max optimization problem, embedded with adversarial tr
aining. These parameters are trained explicitly to achieve improved robustness.
 The extensive results show that our proposed PNI technique effectively improves
 the robustness against a variety of powerful white-box and black-box attacks su
ch as PGD, C&W, FGSM, transferable attack, and ZOO attack. Last but not the leas
t, PNI method improves both clean- and perturbed-data accuracy, in comparison to
 the state-of-the-art defense methods, which outperforms current unbroken PGD de
fense by 1.1% and 6.8% on clean- and perturbed- test data respectively, using Re
sNet-20 architecture.
****

Invariance Matters: Exemplar Memory for Domain Adaptive Person Re-Identification
Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, Yi Yang; Proceedings of the
 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp
. 598-607
This paper considers the domain adaptive person re-identification (re-ID) proble
m: learning a re-ID model from a labeled source domain and an unlabeled target d
omain. Conventional methods are mainly to reduce feature distribution gap betwee
n the source and target domains. However, these studies largely neglect the intr
a-domain variations in the target domain, which contain critical factors influen
cing the testing performance on the target domain. In this work, we comprehensiv
ely investigate into the intra-domain variations of the target domain and propos
e to generalize the re-ID model w.r.t three types of the underlying invariance,
i.e., exemplar-invariance, camera-invariance and neighborhood-invariance. To ach
ieve this goal, an exemplar memory is introduced to store features of the target
 domain and accommodate the three invariance properties. The memory allows us to
 enforce the invariance constraints over global training batch without significa
ntly increasing computation cost. Experiment demonstrates that the three invaria
nce properties and the proposed memory are indispensable towards an effective do
main adaptation system. Results on three re-ID domains show that our domain adap
tation accuracy outperforms the state of the art by a large margin. Code is avai
lable at: https://github.com/zhunzhong07/ECN
****

Dissecting Person Re-Identification From the Viewpoint of Viewpoint
Xiaoxiao Sun, Liang Zheng; Proceedings of the IEEE/CVF Conference on Computer V
ision and Pattern Recognition (CVPR), 2019, pp. 608-617
Variations in visual factors such as viewpoint, pose, illumination and backgroun
d, are usually viewed as important challenges in person re-identification (re-ID
). In spite of acknowledging these factors to be influential, quantitative studi
es on how they affect a re-ID system are still lacking. To derive insights in th
is scientific campaign, this paper makes an early attempt in studying a particul
ar factor, viewpoint. We narrow the viewpoint problem down to the pedestrian rot
ation angle to obtain focused conclusions. In this regard, this paper makes two
contributions to the community. First, we introduce a large-scale synthetic data

engine, PersonX. Composed of hand-crafted 3D person models, the salient characteristic of this engine is "controllable". That is, we are able to synthesize pedestrians by setting the visual variables to arbitrary values. Second, on the 3D data engine, we quantitatively analyze the influence of pedestrian rotation angle on re-ID accuracy. Comprehensively, the person rotation angles are precisely customized from 0 to 360, allowing us to investigate its effect on the training, query, and gallery sets. Extensive experiment helps us have a deeper understanding of the fundamental problems in person re-ID. Our research also provides useful insights for dataset building and future practical usage, e.g., a person of a side view makes a better query.
****

Learning to Reduce Dual-Level Discrepancy for Infrared-Visible Person Re-Identification
Zhixiang Wang, Zheng Wang, Yinqiang Zheng, Yung-Yu Chuang, Shin'ichi Satoh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 618-626
Infrared-Visible person RE-IDentification (IV-REID) is a rising task. Compared to conventional person re-identification (re-ID), IV-REID concerns the additional modality discrepancy originated from the different imaging processes of spectrum cameras, in addition to the person's appearance discrepancy caused by viewpoint changes, pose variations and deformations presented in the conventional re-ID task. The co-existed discrepancies make IV-REID more difficult to solve. Previous methods attempt to reduce the appearance and modality discrepancies simultaneously using feature-level constraints. It is however difficult to eliminate the mixed discrepancies using only feature-level constraints. To address the problem, this paper introduces a novel Dual-level Discrepancy Reduction Learning (D^2RL) scheme which handles the two discrepancies separately. For reducing the modality discrepancy, an image-level sub-network is trained to translate an infrared image into its visible counterpart and a visible image to its infrared version. With the image-level sub-network, we can unify the representations for images with different modalities. With the help of the unified multi-spectral images, a feature-level sub-network is trained to reduce the remaining appearance discrepancy through feature embedding. By cascading the two sub-networks and training them jointly, the dual-level reductions take their responsibilities cooperatively and attentively. Extensive experiments demonstrate the proposed approach outperforms the state-of-the-art methods.
****

Progressive Feature Alignment for Unsupervised Domain Adaptation
Chaoqi Chen, Weiping Xie, Wenbing Huang, Yu Rong, Xinghao Ding, Yue Huang, Tingyang Xu, Junzhou Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 627-636
Unsupervised domain adaptation (UDA) transfers knowledge from a label-rich source domain to a fully-unlabeled target domain. To tackle this task, recent approaches resort to discriminative domain transfer in virtue of pseudo-labels to enforce the class-level distribution alignment across the source and target domains. These methods, however, are vulnerable to the error accumulation and thus incapable of preserving cross-domain category consistency, as the pseudo-labeling accuracy is not guaranteed explicitly. In this paper, we propose the Progressive Feature Alignment Network (PFAN) to align the discriminative features across domains progressively and effectively, via exploiting the intra-class variation in the target domain. To be specific, we first develop an Easy-to-Hard Transfer Strategy (EHTS) and an Adaptive Prototype Alignment (APA) step to train our model iteratively and alternatively. Moreover, upon observing that a good domain adaptation usually requires a non-saturated source classifier, we consider a simple yet efficient way to retard the convergence speed of the source classification loss by further involving a temperature variate into the soft-max function. The extensive experimental results reveal that the proposed PFAN exceeds the state-of-the-art performance on three UDA datasets.
****

Feature-Level Frankenstein: Eliminating Variations for Discriminative Recognitio

n
Xiaofeng Liu, Site Li, Lingsheng Kong, Wanqing Xie, Ping Jia, Jane You, B. V.K. Kumar; Proceedings of the IEEE/CVF Conference on Computer Vision and Patter n Recognition (CVPR), 2019, pp. 637-646
Recent successes of deep learning-based recognition rely on maintaining the cont ent related to the main-task label. However, how to explicitly dispel the noisy signals for better generalization remains an open issue. We systematically summa rize the detrimental factors as task-relevant/irrelevant semantic variations and unspecified latent variation. In this paper, we cast these problems as an adver sarial minimax game in the latent space. Specifically, we propose equipping an e nd-to-end conditional adversarial network with the ability to decompose an input sample into three complementary parts. The discriminative representation inheri ts the desired invariance property guided by prior knowledge of the task, which is marginally independent to the task-relevant/irrelevant semantic and latent va riations. Our proposed framework achieves top performance on a serial of tasks, including digits recognition, lighting, makeup, disguise-tolerant face recogniti on, and facial attributes recognition.
****

Learning a Deep ConvNet for Multi-Label Classification With Partial Labels
Thibaut Durand, Nazanin Mehrasa, Greg Mori; Proceedings of the IEEE/CVF Confer ence on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 647-657
Deep ConvNets have shown great performance for single-label image classification (e.g. ImageNet), but it is necessary to move beyond the single-label classifica tion task because pictures of everyday life are inherently multi-label. Multi-la bel classification is a more difficult task than single-label classification bec ause both the input images and output label spaces are more complex. Furthermore , collecting clean multi-label annotations is more difficult to scale-up than si ngle-label annotations. To reduce the annotation cost, we propose to train a mod el with partial labels i.e. only some labels are known per image. We first empir ically compare different labeling strategies to show the potential for using par tial labels on multi-label datasets. Then to learn with partial labels, we intro duce a new classification loss that exploits the proportion of known labels per example. Our approach allows the use of the same training settings as when learn ing with all the annotations. We further explore several curriculum learning bas ed strategies to predict missing labels. Experiments are performed on three larg e-scale multi-label datasets: MS COCO, NUS-WIDE and Open Images.
****

Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regres sion
Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, Si lvio Savarese; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat tern Recognition (CVPR), 2019, pp. 658-666
Intersection over Union (IoU) is the most popular evaluation metric used in the object detection benchmarks. However, there is a gap between optimizing the comm only used distance losses for regressing the parameters of a bounding box and ma ximizing this metric value. The optimal objective for a metric is the metric its elf. In the case of axis-aligned 2D bounding boxes, it can be shown that IoU can be directly used as a regression loss. However, IoU has a plateau making it inf easible to optimize in the case of non-overlapping bounding boxes. In this paper , we address the this weakness by introducing a generalized version of IoU as bo th a new loss and a new metric. By incorporating this generalized IoU (GIoU) a s a loss into the state-of-the art object detection frameworks, we show a consis tent improvement on their performance using both the standard, IoU based, and ne w, GIoU based, performance measures on popular object detection benchmarks such as PASCAL VOC and MS COCO.
****

Densely Semantically Aligned Person Re-Identification
Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Zhibo Chen; Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 667 -676

We propose a densely semantically aligned person re-identification (re-ID) framework. It fundamentally addresses the body misalignment problem caused by pose/viewpoint variations, imperfect person detection, occlusion, etc.. By leveraging the estimation of the dense semantics of a person image, we construct a set of densely semantically aligned part images (DSAP-images), where the same spatial positions have the same semantics across different person images. We design a two-stream network that consists of a main full image stream (MF-Stream) and a densely semantically-aligned guiding stream (DSAG-Stream). The DSAG-Stream, with the DSAP-images as input, acts as a regulator to guide the MF-Stream to learn densely semantically aligned features from the original image. In the inference, the DSAG-Stream is discarded and only the MF-Stream is needed, which makes the inference system computationally efficient and robust. To our best knowledge, we are the first to make use of fine grained semantics for addressing misalignment problems for re-ID. Our method achieves rank-1 accuracy of 78.9% (new protocol) on the CUHK03 dataset, 90.4% on the CUHK01 dataset, and 95.7% on the Market1501 dataset, outperforming state-of-the-art methods.
****

Generalising Fine-Grained Sketch-Based Image Retrieval
Kaiyue Pang, Ke Li, Yongxin Yang, Honggang Zhang, Timothy M. Hospedales, Tao Xiang, Yi-Zhe Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 677-686
Fine-grained sketch-based image retrieval (FG-SBIR) addresses matching specific photo instance using free-hand sketch as a query modality. Existing models aim to learn an embedding space in which sketch and photo can be directly compared. While successful, they require instance-level pairing within each coarse-grained category as annotated training data. Since the learned embedding space is domain-specific, these models do not generalise well across categories. This limits the practical applicability of FG-SBIR. In this paper, we identify cross-category generalisation for FG-SBIR as a domain generalisation problem, and propose the first solution. Our key contribution is a novel unsupervised learning approach to model a universal manifold of prototypical visual sketch traits. This manifold can then be used to paramaterise the learning of a sketch/photo representation. Model adaptation to novel categories then becomes automatic via embedding the novel sketch in the manifold and updating the representation and retrieval function accordingly. Experiments on the two largest FG-SBIR datasets, Sketchy and QMUL-Shoe-V2, demonstrate the efficacy of our approach in enabling cross-category generalisation of FG-SBIR.
****

Adapting Object Detectors via Selective Cross-Domain Alignment
Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, Dahua Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 687-696
State-of-the-art object detectors are usually trained on public datasets. They often face substantial difficulties when applied to a different domain, where the imaging condition differs significantly and the corresponding annotated data are unavailable (or expensive to acquire). A natural remedy is to adapt the model by aligning the image representations on both domains. This can be achieved, for example, by adversarial learning, and has been shown to be effective in tasks like image classification. However, we found that in object detection, the improvement obtained in this way is quite limited. An important reason is that conventional domain adaptation methods strive to align images as a whole, while object detection, by nature, focuses on local regions that may contain objects of interest. Motivated by this, we propose a novel approach to domain adaption for object detection to handle the issues in "where to look" and "how to align". Our key idea is to mine the discriminative regions, namely those that are directly pertinent to object detection, and focus on aligning them across both domains. Experiments show that the proposed method performs remarkably better than existing methods with about 4% 6% improvement under various domain-shift scenarios while keeping good scalability.
****

Cyclic Guidance for Weakly Supervised Joint Detection and Segmentation
Yunhang Shen, Rongrong Ji, Yan Wang, Yongjian Wu, Liujuan Cao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 697-707
Weakly supervised learning has attracted growing research attention due to the significant saving in annotation cost for tasks that require intra-image annotations, such as object detection and semantic segmentation. To this end, existing weakly supervised object detection and semantic segmentation approaches follow an iterative label mining and model training pipeline. However, such a self-enforcement pipeline makes both tasks easy to be trapped in local minimums. In this paper, we join weakly supervised object detection and segmentation tasks with a multi-task learning scheme for the first time, which uses their respective failure patterns to complement each other's learning. Such cross-task enforcement helps both tasks to leap out of their respective local minimums. In particular, we present an efficient and effective framework termed Weakly Supervised Joint Detection and Segmentation (WS-JDS). WS-JDS has two branches for the above two tasks, which share the same backbone network. In the learning stage, it uses the same cyclic training paradigm but with a specific loss function such that the two branches benefit each other. Extensive experiments have been conducted on the widely-used Pascal VOC and COCO benchmarks, which demonstrate that our model has achieved competitive performance with the state-of-the-art algorithms.
****
Thinking Outside the Pool: Active Training Image Creation for Relative Attributes
Aron Yu, Kristen Grauman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 708-718
Current wisdom suggests more labeled image data is always better, and obtaining labels is the bottleneck. Yet curating a pool of sufficiently diverse and informative images is itself a challenge. In particular, training image curation is problematic for fine-grained attributes, where the subtle visual differences of interest may be rare within traditional image sources. We propose an active image generation approach to address this issue. The main idea is to jointly learn the attribute ranking task while also learning to generate novel realistic image samples that will benefit that task. We introduce an end-to-end framework that dynamically "imagines" image pairs that would confuse the current model, presents them to human annotators for labeling, then improves the predictive model with the new examples. On two datasets, we show that by thinking outside the pool of real images, our approach gains generalization accuracy on challenging fine-grained attribute comparisons.
****
Generalizable Person Re-Identification by Domain-Invariant Mapping Network
Jifei Song, Yongxin Yang, Yi-Zhe Song, Tao Xiang, Timothy M. Hospedales; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 719-728
We aim to learn a domain generalizable person re-identification (ReID) model. When such a model is trained on a set of source domains (ReID datasets collected from different camera networks), it can be directly applied to any new unseen dataset for effective ReID without any model updating. Despite its practical value in real-world deployments, generalizable ReID has seldom been studied. In this work, a novel deep ReID model termed Domain-Invariant Mapping Network (DIMN) is proposed. DIMN is designed to learn a mapping between a person image and its identity classifier, i.e., it produces a classifier using a single shot. To make the model domain-invariant, we follow a meta-learning pipeline and sample a subset of source domain training tasks during each training episode. However, the model is significantly different from conventional meta-learning methods in that: (1) no model updating is required for the target domain, (2) different training tasks share a memory bank for maintaining both scalability and discrimination ability, and (3) it can be used to match an arbitrary number of identities in a target domain. Extensive experiments on a newly proposed large-scale ReID domain generalization benchmark show that our DIMN significantly outperforms alternativ

e domain generalization or meta-learning methods.
****

Visual Attention Consistency Under Image Transforms for Multi-Label Image Classification
Hao Guo, Kang Zheng, Xiaochuan Fan, Hongkai Yu, Song Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 729-739
Human visual perception shows good consistency for many multi-label image classification tasks under certain spatial transforms, such as scaling, rotation, flipping and translation. This has motivated the data augmentation strategy widely used in CNN classifier training -- transformed images are included for training by assuming the same class labels as their original images. In this paper, we further propose the assumption of perceptual consistency of visual attention regions for classification under such transforms, i.e., the attention region for a classification follows the same transform if the input image is spatially transformed. While the attention regions of CNN classifiers can be derived as an attention heatmap in middle layers of the network, we find that their consistency under many transforms are not preserved. To address this problem, we propose a two-branch network with an original image and its transformed image as inputs and introduce a new attention consistency loss that measures the attention heatmap consistency between two branches. This new loss is then combined with multi-label image classification loss for network training. Experiments on three datasets verify the superiority of the proposed network by achieving new state-of-the-art classification performance.
****

Re-Ranking via Metric Fusion for Object Retrieval and Person Re-Identification
Song Bai, Peng Tang, Philip H.S. Torr, Longin Jan Latecki; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 740-749
This work studies the unsupervised re-ranking procedure for object retrieval and person re-identification with a specific concentration on an ensemble of multiple metrics (or similarities). While the re-ranking step is involved by running a diffusion process on the underlying data manifolds, the fusion step can leverage the complementarity of multiple metrics. We give a comprehensive summary of existing fusion with diffusion strategies, and systematically analyze their pros and cons. Based on the analysis, we propose a unified yet robust algorithm which inherits their advantages and discards their disadvantages. Hence, we call it Unified Ensemble Diffusion (UED). More interestingly, we derive that the inherited properties indeed stem from a theoretical framework, where the relevant works can be elegantly summarized as special cases of UED by imposing additional constraints on the objective function and varying the solver of similarity propagation. Extensive experiments with 3D shape retrieval, image retrieval and person re-identification demonstrate that the proposed framework outperforms the state of the arts, and at the same time suggest that re-ranking via metric fusion is a promising tool to further improve the retrieval performance of existing algorithms.
****

Unsupervised Open Domain Recognition by Semantic Discrepancy Minimization
Junbao Zhuo, Shuhui Wang, Shuhao Cui, Qingming Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 750-759
We address the unsupervised open domain recognition (UODR) problem, where categories in labeled source domain S is only a subset of those in unlabeled target domain T. The task is to correctly classify all samples in T including known and unknown categories. UODR is challenging due to the domain discrepancy, which becomes even harder to bridge when a large number of unknown categories exist in T. Moreover, the classification rules propagated by graph CNN (GCN) may be distracted by unknown categories and lack generalization capability. To measure the domain discrepancy for asymmetric label space between S and T, we propose Semantic-Guided Matching Discrepancy (SGMD), which first employs instance matching

between S and T, and then the discrepancy is measured by a weighted feature distance between matched instances. We further design a limited balance constraint to achieve a more balanced classification output on known and unknown categories. We develop Unsupervised Open Domain Transfer Network (UODTN), which learns both the backbone classification network and GCN jointly by reducing the SGMD, enforcing the limited balance constraint and minimizing the classification loss on S. UODTN better preserves the semantic structure and enforces the consistency between the learned domain invariant visual features and the semantic embeddings. Experimental results show superiority of our method on recognizing images of both known and unknown categories.
****

Weakly Supervised Person Re-Identification
Jingke Meng, Sheng Wu, Wei-Shi Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 760-769
In the conventional person re-id setting, it is assumed that the labeled images are the person images within the bounding box for each individual; this labeling across multiple nonoverlapping camera views from raw video surveillance is costly and time-consuming. To overcome this difficulty, we consider weakly supervised person re-id modeling. The weak setting refers to matching a target person with an untrimmed gallery video where we only know that the identity appears in the video without the requirement of annotating the identity in any frame of the video during the training procedure. Hence, for a video, there could be multiple video-level labels. We cast this weakly supervised person re-id challenge into a multi-instance multi-label learning (MIML) problem. In particular, we develop a Cross-View MIML (CV-MIML) method that is able to explore potential intraclass person images from all the camera views by incorporating the intra-bag alignment and the cross-view bag alignment. Finally, the CV-MIML method is embedded into an existing deep neural network for developing the Deep Cross-View MIML (Deep CV-MIML) model. We have performed extensive experiments to show the feasibility of the proposed weakly supervised setting and verify the effectiveness of our method compared to related methods on four weakly labeled datasets.
****

PointRCNN: 3D Object Proposal Generation and Detection From Point Cloud
Shaoshuai Shi, Xiaogang Wang, Hongsheng Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 770-779
In this paper, we propose PointRCNN for 3D object detection from raw point cloud. The whole framework is composed of two stages: stage-1 for the bottom-up 3D proposal generation and stage-2 for refining proposals in the canonical coordinates to obtain the final detection results. Instead of generating proposals from RGB image or projecting point cloud to bird's view or voxels as previous methods do, our stage-1 sub-network directly generates a small number of high-quality 3D proposals from point cloud in a bottom-up manner via segmenting the point cloud of the whole scene into foreground points and background. The stage-2 sub-network transforms the pooled points of each proposal to canonical coordinates to learn better local spatial features, which is combined with global semantic features of each point learned in stage-1 for accurate box refinement and confidence prediction. Extensive experiments on the 3D detection benchmark of KITTI dataset show that our proposed architecture outperforms state-of-the-art methods with remarkable margins by using only point cloud as input. The code is available at https://github.com/sshaoshuai/PointRCNN.
****

Automatic Adaptation of Object Detectors to New Domains Using Self-Training
Aruni RoyChowdhury, Prithvijit Chakrabarty, Ashish Singh, SouYoung Jin, Huaizu Jiang, Liangliang Cao, Erik Learned-Miller; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 780-790
This work addresses the unsupervised adaptation of an existing object detector to a new target domain. We assume that a large number of unlabeled videos from this domain are readily available. We automatically obtain labels on the target data by using high-confidence detections from the existing detector, augmented with hard (misclassified) examples acquired by exploiting temporal cues using a tra

cker. These automatically-obtained labels are then used for re-training the orig inal model. A modified knowledge distillation loss is proposed, and we investiga te several ways of assigning soft-labels to the training examples from the targe t domain. Our approach is empirically evaluated on challenging face and pedestri an detection tasks: a face detector trained on WIDER-Face, which consists of hig h-quality images crawled from the web, is adapted to a large-scale surveillance data set; a pedestrian detector trained on clear, daytime images from the BDD-1 00K driving data set is adapted to all other scenarios such as rainy, foggy, nig ht-time. Our results demonstrate the usefulness of incorporating hard examples o btained from tracking, the advantage of using soft-labels via distillation loss versus hard-labels, and show promising performance as a simple method for unsupe rvised domain adaptation of object detectors, with minimal dependence on hyper-p arameters.

****

Deep Sketch-Shape Hashing With Segmented 3D Stochastic Viewing

Jiaxin Chen, Jie Qin, Li Liu, Fan Zhu, Fumin Shen, Jin Xie, Ling Shao; Pro ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 791-800

Sketch-based 3D shape retrieval has been extensively studied in recent works, mo st of which focus on improving the retrieval accuracy, whilst neglecting the eff iciency. In this paper, we propose a novel framework for efficient sketch-based 3D shape retrieval, i.e., Deep Sketch-Shape Hashing (DSSH), which tackles the ch allenging problem from two perspectives. Firstly, we propose an intuitive 3D sha pe representation method to deal with unaligned shapes with arbitrary poses. Spe cifically, the proposed Segmented Stochastic-viewing Shape Network models discri minative 3D representations by a set of 2D images rendered from multiple views, which are stochastically selected from non-overlapping spatial segments of a 3D sphere. Secondly, Batch-Hard Binary Coding (BHBC) is developed to learn semantic s-preserving compact binary codes by mining the hardest samples. The overall fra mework is jointly learned by developing an alternating iteration algorithm. Exte nsive experimental results on three benchmarks show that DSSH improves both the retrieval efficiency and accuracy remarkably, compared to the state-of-the-art m ethods.

****

Generative Dual Adversarial Network for Generalized Zero-Shot Learning

He Huang, Changhu Wang, Philip S. Yu, Chang-Dong Wang; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 80 1-810

This paper studies the problem of generalized zero-shot learning which requires the model to train on image-label pairs from some seen classes and test on the t ask of classifying new images from both seen and unseen classes. In this paper, we propose a novel model that provides a unified framework for three different approaches: visual->semantic mapping, semantic->visual mapping, and metric learn ing. Specifically, our proposed model consists of a feature generator that can g enerate various visual features given class embeddings as input, a regressor tha t maps each visual feature back to its corresponding class embedding, and a disc riminator that learns to evaluate the closeness of an image feature and a class embedding. All three components are trained under the combination of cyclic cons istency loss and dual adversarial loss. Experimental results show that our model not only preserves higher accuracy in classifying images from seen classes, but also performs better than existing state-of-the-art models in in classifying im ages from unseen classes.

****

Query-Guided End-To-End Person Search

Bharti Munjal, Sikandar Amin, Federico Tombari, Fabio Galasso; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019 , pp. 811-820

Person search has recently gained attention as the novel task of finding a perso n, provided as a cropped sample, from a gallery of non-cropped images, whereby s everal other people are also visible. We believe that i. person detection and re

-identification should be pursued in a joint optimization framework and that ii. the person search should leverage the query image extensively (e.g. emphasizing unique query patterns). However, so far, no prior art realizes this. We introduce a novel query-guided end-to-end person search network (QEEPS) to address both aspects. We leverage a most recent joint detector and re-identification work, OIM [37]. We extend this with i. a query-guided Siamese squeeze-and-excitation network (QSSE-Net) that uses global context from both the query and gallery images, ii. a query-guided region proposal network (QRPN) to produce query-relevant proposals, and iii. a query-guided similarity subnetwork (QSimNet), to learn a query-guided re-identification score. QEEPS is the first end-to-end query-guided detection and re-id network. On both the most recent CUHK-SYSU [37] and PRW [46] datasets, we outperform the previous state-of-the-art by a large margin.

****

Libra R-CNN: Towards Balanced Learning for Object Detection

Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, Dahua Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 821-830

Compared with model architectures, the training process, which is also crucial to the success of detectors, has received relatively less attention in object detection. In this work, we carefully revisit the standard training practice of detectors, and find that the detection performance is often limited by the imbalance during the training process, which generally consists in three levels - sample level, feature level, and objective level. To mitigate the adverse effects caused thereby, we propose Libra R-CNN, a simple but effective framework towards balanced learning for object detection. It integrates three novel components: IoU-balanced sampling, balanced feature pyramid, and balanced L1 loss, respectively for reducing the imbalance at sample, feature, and objective level. Benefitted from the overall balanced design, Libra R-CNN significantly improves the detection performance. Without bells and whistles, it achieves 2.5 points and 2.0 points higher Average Precision (AP) than FPN Faster R-CNN and RetinaNet respectively on MSCOCO.

****

Learning a Unified Classifier Incrementally via Rebalancing

Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, Dahua Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 831-839

Conventionally, deep neural networks are trained offline, relying on a large dataset prepared in advance. This paradigm is often challenged in real-world applications, e.g. online services that involve continuous streams of incoming data. Recently, incremental learning receives increasing attention, and is considered as a promising solution to the practical challenges mentioned above. However, it has been observed that incremental learning is subject to a fundamental difficulty -- catastrophic forgetting, namely adapting a model to new data often results in severe performance degradation on previous tasks or classes. Our study reveals that the imbalance between previous and new data is a crucial cause to this problem. In this work, we develop a new framework for incrementally learning a unified classifier, e.g. a classifier that treats both old and new classes uniformly. Specifically, we incorporate three components, cosine normalization, less-forget constraint, and inter-class separation, to mitigate the adverse effects of the imbalance. Experiments show that the proposed method can effectively rebalance the training process, thus obtaining superior performance compared to the existing methods. On CIFAR-100 and ImageNet, our method can reduce the classification errors by more than 6% and 13% respectively, under the incremental setting of 10 phases.

****

Feature Selective Anchor-Free Module for Single-Shot Object Detection

Chenchen Zhu, Yihui He, Marios Savvides; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 840-849

We motivate and present feature selective anchor-free (FSAF) module, a simple and effective building block for single-shot object detectors. It can be plugged i

nto single-shot detectors with feature pyramid structure. The FSAF module addres
ses two limitations brought up by the conventional anchor-based detection: 1) he
uristic-guided feature selection; 2) overlap-based anchor sampling. The general
concept of the FSAF module is online feature selection applied to the training o
f multi-level anchor-free branches. Specifically, an anchor-free branch is attac
hed to each level of the feature pyramid, allowing box encoding and decoding in
the anchor-free manner at an arbitrary level. During training, we dynamically as
sign each instance to the most suitable feature level. At the time of inference,
 the FSAF module can work independently or jointly with anchor-based branches. W
e instantiate this concept with simple implementations of anchor-free branches a
nd online feature selection strategy. Experimental results on the COCO detection
 track show that our FSAF module performs better than anchor-based counterparts
while being faster. When working jointly with anchor-based branches, the FSAF mo
dule robustly improves the baseline RetinaNet by a large margin under various se
ttings, while introducing nearly free inference overhead. And the resulting best
 model can achieve a state-of-the-art 44.6% mAP, outperforming all existing sing
le-shot detectors on COCO.
****

Bottom-Up Object Detection by Grouping Extreme and Center Points
Xingyi Zhou, Jiacheng Zhuo, Philipp Krahenbuhl; Proceedings of the IEEE/CVF Co
nference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 850-859
With the advent of deep learning, object detection drifted from a bottom-up to a
 top-down recognition problem. State of the art algorithms enumerate a near-exha
ustive list of object locations and classify each into: object or not. In this p
aper, we show that bottom-up approaches still perform competitively. We detect f
our extreme points (top-most, left-most, bottom-most, right-most) and one center
 point of objects using a standard keypoint estimation network. We group the fiv
e keypoints into a bounding box if they are geometrically aligned. Object detect
ion is then a purely appearance-based keypoint estimation problem, without regio
n classification or implicit feature learning. The proposed method performs on-p
ar with the state-of-the-art region based detection methods, with a bounding box
 AP of 43.7% on COCO test-dev. In addition, our estimated extreme points directl
y span a coarse octagonal mask, with a COCO Mask AP of 18.9%, much better than t
he Mask AP of vanilla bounding boxes. Extreme point guided segmentation further
improves this to 34.6% Mask AP.
****

Feature Distillation: DNN-Oriented JPEG Compression Against Adversarial Examples
Zihao Liu, Qi Liu, Tao Liu, Nuo Xu, Xue Lin, Yanzhi Wang, Wujie Wen; Proce
edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV
PR), 2019, pp. 860-868
Image compression-based approaches for defending against the adversarial-example
 attacks, which threaten the safety use of deep neural networks (DNN), have been
 investigated recently. However, prior works mainly rely on directly tuning para
meters like compression rate, to blindly reduce image features, thereby lacking
guarantee on both defense efficiency (i.e. accuracy of polluted images) and clas
sification accuracy of benign images, after applying defense methods. To overcom
e these limitations, we propose a JPEG-based defensive compression framework, na
mely "feature distillation", to effectively rectify adversarial examples without
 impacting classification accuracy on benign data. Our framework significantly e
scalates the defense efficiency with marginal accuracy reduction using a twostep
 method: First, we maximize malicious features filtering of adversarial input pe
rturbations by developing defensive quantization in frequency domain of JPEG com
pression or decompression, guided by a semi-analytical method; Second, we suppre
ss the distortions of benign features to restore classification accuracy through
 a DNN-oriented quantization refine process. Our experimental results show that
proposed "feature distillation" can significantly surpass the latest input-trans
formation based mitigations such as Quilting and TV Minimization in three aspect
s, including defense efficiency (improve classification accuracy from 20% to
 90% on adversarial examples), accuracy of benign images after defense (<= 1% ac
curacy degradation), and processing time per image (259x Speedup). Moreover, o

ur solution also can provide the best defense efficiency (60% accuracy) agains t the latest BPDA attack with least accuracy reduction (1%) on benign images a mong all other input-transformation based defense methods.
****

SCOPS: Self-Supervised Co-Part Segmentation
Wei-Chih Hung, Varun Jampani, Sifei Liu, Pavlo Molchanov, Ming-Hsuan Yang, Jan Kautz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 869-878
Parts provide a good intermediate representation of objects that is robust with respect to camera, pose and appearance variations. Existing work on part segment ation is dominated by supervised approaches that rely on large amounts of manual annotations and also can not generalize to unseen object categories. We propose a self-supervised deep learning approach for part segmentation, where we devise several loss functions that aids in predicting part segments that are geometric ally concentrated, robust to object variations and are also semantically consist ent across different object instances. Extensive experiments on different types of image collections demonstrate that our approach can produce part segments tha t adhere to object boundaries and also more semantically consistent across objec t instances compared to existing self-supervised techniques.
****

Unsupervised Moving Object Detection via Contextual Information Separation
Yanchao Yang, Antonio Loquercio, Davide Scaramuzza, Stefano Soatto; Proceedin gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 879-888
We propose an adversarial contextual model for detecting moving objects in image s. A deep neural network is trained to predict the optical flow in a region usin g information from everywhere else but that region (context), while another netw ork attempts to make such context as uninformative as possible. The result is a model where hypotheses naturally compete with no need for explicit regularizatio n or hyper-parameter tuning. Although our method requires no supervision whatsoe ver, it outperforms several methods that are pre-trained on large annotated data sets. Our model can be thought of as a generalization of classical variational g enerative region-based segmentation, but in a way that avoids explicit regulariz ation or solution of partial differential equations at run-time.
****

Pose2Seg: Detection Free Human Instance Segmentation
Song-Hai Zhang, Ruilong Li, Xin Dong, Paul Rosin, Zixi Cai, Xi Han, Dingch eng Yang, Haozhi Huang, Shi-Min Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 889-898
The standard approach to image instance segmentation is to perform the object de tection first, and then segment the object from the detection bounding-box. More recently, deep learning methods like Mask R-CNN perform them jointly. However, little research takes into account the uniqueness of the "human" category, which can be well defined by the pose skeleton. Moreover, the human pose skeleton can be used to better distinguish instances with heavy occlusion than using boundin g-boxes. In this paper, we present a brand new pose-based instance segmentation framework for humans which separates instances based on human pose, rather than proposal region detection. We demonstrate that our pose-based framework can achi eve better accuracy than the state-of-art detection-based approach on the human instance segmentation problem, and can moreover better handle occlusion. Further more, there are few public datasets containing many heavily occluded humans alon g with comprehensive annotations, which makes this a challenging problem seldom noticed by researchers. Therefore, in this paper we introduce a new benchmark "O ccluded Human (OCHuman)", which focuses on occluded humans with comprehensive an notations including bounding-box, human pose and instance masks. This dataset c ontains 8110 detailed annotated human instances within 4731 images. With an aver age 0.67 MaxIoU for each person, OCHuman is the most complex and challenging dat aset related to human instance segmentation. Through this dataset, we want to em phasize occlusion as a challenging problem for researchers to study.
****

DrivingStereo: A Large-Scale Dataset for Stereo Matching in Autonomous Driving Scenarios
Guorun Yang, Xiao Song, Chaoqin Huang, Zhidong Deng, Jianping Shi, Bolei Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 899-908
Great progress has been made on estimating disparity maps from stereo images. However, with the limited stereo data available in the existing datasets and unstable ranging precision of current stereo methods, industry-level stereo matching in autonomous driving remains challenging. In this paper, we construct a novel large-scale stereo dataset named DrivingStereo. It contains over 180k images covering a diverse set of driving scenarios, which is hundreds of times larger than the KITTI Stereo dataset. High-quality labels of disparity are produced by a model-guided filtering strategy from multi-frame LiDAR points. For better evaluations, we present two new metrics for stereo matching in the driving scenes, i.e. a distance-aware metric and a semantic-aware metric. Extensive experiments show that compared with the models trained on FlyingThings3D or Cityscapes, the models trained on our DrivingStereo achieve higher generalization accuracy in real-world driving scenes, while the proposed metrics better evaluate the stereo methods on all-range distances and across different classes. Our dataset and code are available at https://drivingstereo-dataset.github.io.
****

PartNet: A Large-Scale Benchmark for Fine-Grained and Hierarchical Part-Level 3D Object Understanding
Kaichun Mo, Shilin Zhu, Angel X. Chang, Li Yi, Subarna Tripathi, Leonidas J. Guibas, Hao Su; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 909-918
We present PartNet: a consistent, large-scale dataset of 3D objects annotated with fine-grained, instance-level, and hierarchical 3D part information. Our dataset consists of 573,585 part instances over 26,671 3D models covering 24 object categories. This dataset enables and serves as a catalyst for many tasks such as shape analysis, dynamic 3D scene modeling and simulation, affordance analysis, and others. Using our dataset, we establish three benchmarking tasks for evaluating 3D part recognition: fine-grained semantic segmentation, hierarchical semantic segmentation, and instance segmentation. We benchmark four state-of-the-art 3D deep learning algorithms for fine-grained semantic segmentation and three baseline methods for hierarchical semantic segmentation. We also propose a baseline method for part instance segmentation and demonstrate its superior performance over existing methods.
****

A Dataset and Benchmark for Large-Scale Multi-Modal Face Anti-Spoofing
Shifeng Zhang, Xiaobo Wang, Ajian Liu, Chenxu Zhao, Jun Wan, Sergio Escalera, Hailin Shi, Zezheng Wang, Stan Z. Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 919-928
Face anti-spoofing is essential to prevent face recognition systems from a security breach. Much of the progresses have been made by the availability of face anti-spoofing benchmark datasets in recent years. However, existing face anti-spoofing benchmarks have limited number of subjects (<=170) and modalities (<=2), which hinder the further development of the academic community. To facilitate face anti-spoofing research, we introduce a large-scale multi-modal dataset, namely CASIA-SURF, which is the largest publicly available dataset for face anti-spoofing in terms of both subjects and visual modalities. Specifically, it consists of 1,000 subjects with 21,000 videos and each sample has 3 modalities (i.e., RGB, Depth and IR). We also provide a measurement set, evaluation protocol and training/validation/testing subsets, developing a new benchmark for face anti-spoofing. Moreover, we present a new multi-modal fusion method as baseline, which performs feature re-weighting to select the more informative channel features while suppressing the less useful ones for each modal. Extensive experiments have been conducted on the proposed dataset to verify its significance and generalization capability. The dataset is available at https://sites.google.com/qq.com/chalearnfacespoofingattackdete/.

****
Unsupervised Learning of Consensus Maximization for 3D Vision Problems
Thomas Probst, Danda Pani Paudel, Ajad Chhatkuli, Luc Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 929-938
Consensus maximization is a key strategy in 3D vision for robust geometric model estimation from measurements with outliers. Generic methods for consensus maximization, such as Random Sampling and Consensus (RANSAC), have played a tremendous role in the success of 3D vision, in spite of the ubiquity of outliers. However, replicating the same generic behaviour in a deeply learned architecture, using supervised approaches, has proven to be difficult. In that context, unsupervised methods have a huge potential to adapt to any unseen data distribution, and therefore are highly desirable. In this paper, we propose for the first time an unsupervised learning framework for consensus maximization, in the context of solving 3D vision problems. For that purpose, we establish a relationship between inlier measurements, represented by an ideal of inlier set, and the subspace of polynomials representing the space of target transformations. Using this relationship, we derive a constraint that must be satisfied by the sought inlier set. This constraint can be tested without knowing the transformation parameters, therefore allows us to efficiently define the geometric model fitting cost. This model fitting cost is used as a supervisory signal for learning consensus maximization, where the learning process seeks for the largest measurement set that minimizes the proposed model fitting cost. Using our method, we solve a diverse set of 3D vision problems, including 3D-3D matching, non-rigid 3D shape matching with piece-wise rigidity and image-to-image matching. Despite being unsupervised, our method outperforms RANSAC in all three tasks for several datasets.
****
VizWiz-Priv: A Dataset for Recognizing the Presence and Purpose of Private Visual Information in Images Taken by Blind People
Danna Gurari, Qing Li, Chi Lin, Yinan Zhao, Anhong Guo, Abigale Stangl, Jeffrey P. Bigham; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 939-948
We introduce the first visual privacy dataset originating from people who are blind in order to better understand their privacy disclosures and to encourage the development of algorithms that can assist in preventing their unintended disclosures. It includes 8,862 regions showing private content across 5,537 images taken by blind people. Of these, 1,403 are paired with questions and 62% of those directly ask about the private content. Experiments demonstrate the utility of this data for predicting whether an image shows private information and whether a question asks about the private content in an image. The dataset is publicly-shared at http://vizwiz.org/data/.
****
Structural Relational Reasoning of Point Clouds
Yueqi Duan, Yu Zheng, Jiwen Lu, Jie Zhou, Qi Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 949-958
The symmetry for the corners of a box, the continuity for the surfaces of a monitor, the linkage between the torso and other body parts --- it suggests that 3D objects may have common and underlying inner relations between local structures, and it is a fundamental ability for intelligent species to reason for them. In this paper, we propose an effective plug-and-play module called the structural relation network (SRN) to reason about the structural dependencies of local regions in 3D point clouds. Existing network architectures on point sets such as PointNet++ capture local structures individually, without considering their inner interactions. Instead, our SRN simultaneously exploits local information by modeling their geometrical and locational relations, which play critical roles for our humans to understand 3D objects. The proposed SRN module is simple, interpretable, and does not require any additional supervision signals, which can be easily equipped with the existing networks. Experimental results on benchmark datasets indicate promising boosts on the tasks of 3D point cloud classification and seg

mentation by capturing structural relations with the SRN module.
****

MVF-Net: Multi-View 3D Face Morphable Model Regression
Fanzi Wu, Linchao Bao, Yajing Chen, Yonggen Ling, Yibing Song, Songnan Li, King Ngi Ngan, Wei Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 959-968
We address the problem of recovering the 3D geometry of a human face from a set of facial images in multiple views. While recent studies have shown impressive progress in 3D Morphable Model (3DMM) based facial reconstruction, the settings are mostly restricted to a single view. There is an inherent drawback in the single-view setting: the lack of reliable 3D constraints can cause unresolvable ambiguities. We in this paper explore 3DMM-based shape recovery in a different setting, where a set of multi-view facial images are given as input. A novel approach is proposed to regress 3DMM parameters from multi-view inputs with an end-to-end trainable Convolutional Neural Network (CNN). Multi-view geometric constraints are incorporated into the network by establishing dense correspondences between different views leveraging a novel self-supervised view alignment loss. The main ingredient of the view alignment loss is a differentiable dense optical flow estimator that can backpropagate the alignment errors between an input view and a synthetic rendering from another input view, which is projected to the target view through the 3D shape to be inferred. Through minimizing the view alignment loss, better 3D shapes can be recovered such that the synthetic projections from one view to another can better align with the observed image. Extensive experiments demonstrate the superiority of the proposed method over other 3DMM methods.
****

Photometric Mesh Optimization for Video-Aligned 3D Object Reconstruction
Chen-Hsuan Lin, Oliver Wang, Bryan C. Russell, Eli Shechtman, Vladimir G. Kim, Matthew Fisher, Simon Lucey; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 969-978
In this paper, we address the problem of 3D object mesh reconstruction from RGB videos. Our approach combines the best of multi-view geometric and data-driven methods for 3D reconstruction by optimizing object meshes for multi-view photometric consistency while constraining mesh deformations with a shape prior. We pose this as a piecewise image alignment problem for each mesh face projection. Our approach allows us to update shape parameters from the photometric error without any depth or mask information. Moreover, we show how to avoid a degeneracy of zero photometric gradients via rasterizing from a virtual viewpoint. We demonstrate 3D object mesh reconstruction results from both synthetic and real-world videos with our photometric mesh optimization, which is unachievable with either naive mesh generation networks or traditional pipelines of surface reconstruction without heavy manual post-processing.
****

Guided Stereo Matching
Matteo Poggi, Davide Pallotti, Fabio Tosi, Stefano Mattoccia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 979-988
Stereo is a prominent technique to infer dense depth maps from images, and deep learning further pushed forward the state-of-the-art, making end-to-end architectures unrivaled when enough data is available for training. However, deep networks suffer from significant drops in accuracy when dealing with new environments. Therefore, in this paper, we introduce Guided Stereo Matching, a novel paradigm leveraging a small amount of sparse, yet reliable depth measurements retrieved from an external source enabling to ameliorate this weakness. The additional sparse cues required by our method can be obtained with any strategy (e.g., a LiDAR) and used to enhance features linked to corresponding disparity hypotheses. Our formulation is general and fully differentiable, thus enabling to exploit the additional sparse inputs in pre-trained deep stereo networks as well as for training a new instance from scratch. Extensive experiments on three standard datasets and two state-of-the-art deep architectures show that even with a small set of sparse input cues, i) the proposed paradigm enables significant improvements to

pre-trained networks. Moreover, ii) training from scratch notably increases accuracy and robustness to domain shifts. Finally, iii) it is suited and effective even with traditional stereo algorithms such as SGM.
****

Unsupervised Event-Based Learning of Optical Flow, Depth, and Egomotion
Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, Kostas Daniilidis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 989-997
In this work, we propose a novel framework for unsupervised learning for event cameras that learns motion information from only the event stream. In particular, we propose an input representation of the events in the form of a discretized volume that maintains the temporal distribution of the events, which we pass through a neural network to predict the motion of the events. This motion is used to attempt to remove any motion blur in the event image. We then propose a loss function applied to the motion compensated event image that measures the motion blur in this image. We train two networks with this framework, one to predict optical flow, and one to predict egomotion and depths, and evaluate these networks on the Multi Vehicle Stereo Event Camera dataset, along with qualitative results from a variety of different scenes.
****

Modeling Local Geometric Structure of 3D Point Clouds Using Geo-CNN
Shiyi Lan, Ruichi Yu, Gang Yu, Larry S. Davis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 998-1008
Recent advances in deep convolutional neural networks (CNNs) have motivated researchers to adapt CNNs to directly model points in 3D point clouds. Modeling local structure has been proven to be important for the success of convolutional architectures, and researchers exploited the modeling of local point sets in the feature extraction hierarchy. However, limited attention has been paid to explicitly model the geometric structure amongst points in a local region. To address this problem, we propose Geo-CNN, which applies a generic convolution-like operation dubbed as GeoConv to each point and its local neighborhood. Local geometric relationships among points are captured when extracting edge features between the center and its neighboring points. We first decompose the edge feature extraction process onto three orthogonal bases, and then aggregate the extracted features based on the angles between the edge vector and the bases. This encourages the network to preserve the geometric structure in Euclidean space throughout the feature extraction hierarchy. GeoConv is a generic and efficient operation that can be easily integrated into 3D point cloud analysis pipelines for multiple applications. We evaluate Geo-CNN on ModelNet40 and KITTI and achieve state-of-the-art performance.
****

3D Point Capsule Networks
Yongheng Zhao, Tolga Birdal, Haowen Deng, Federico Tombari; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1009-1018
In this paper, we propose 3D point-capsule networks, an auto-encoder designed to process sparse 3D point clouds while preserving spatial arrangements of the input data. 3D capsule networks arise as a direct consequence of our unified formulation of the common 3D auto-encoders. The dynamic routing scheme and the peculiar 2D latent space deployed by our capsule networks bring in improvements for several common point cloud-related tasks, such as object classification, object reconstruction and part segmentation as substantiated by our extensive evaluations. Moreover, it enables new applications such as part interpolation and replacement.
****

GS3D: An Efficient 3D Object Detection Framework for Autonomous Driving
Buyu Li, Wanli Ouyang, Lu Sheng, Xingyu Zeng, Xiaogang Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1019-1028
We present an efficient 3D object detection framework based on a single RGB imag

e in the scenario of autonomous driving. Our efforts are put on extracting the u
nderlying 3D information in a 2D image and determining the accurate 3D bounding
box of object without point cloud or stereo data. Leveraging the off-the-shelf 2
D object detector, we propose an artful approach to efficiently obtain a coarse
cuboid for each predicted 2D box. The coarse cuboid has enough accuracy to guide
 us to determine the 3D box of the object by refinement. In contrast to previous
 state-of-the-art methods that only use the features extracted from the 2D bound
ing box for box refinement, we explore the 3D structure information of the objec
t by employing the visual features of visible surfaces. The new features from su
rfaces are utilized to eliminate the problem of representation ambiguity brought
 by only using 2D bounding box. Moreover, we investigate different methods of 3D
 box refinement and discover that a classification formulation with quality awar
e loss have much better performance than regression. Evaluated on KITTI benchmar
k, our approach outperforms current state-of-the-art methods for single RGB imag
e based 3D object detection.
****

Single-Image Piece-Wise Planar 3D Reconstruction via Associative Embedding
Zehao Yu, Jia Zheng, Dongze Lian, Zihan Zhou, Shenghua Gao; Proceedings of t
he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019,
pp. 1029-1037
Single-image piece-wise planar 3D reconstruction aims to simultaneously segment
 plane instances and recover 3D plane parameters from an image. Most recent appro
aches leverage convolutional neural networks (CNNs) and achieve promising result
s. However, these methods are limited to detecting a fixed number of planes with
 certain learned order. To tackle this problem, we propose a novel two-stage met
hod based on associative embedding, inspired by its recent success in instance s
egmentation. In the first stage, we train a CNN to map each pixel to an embeddin
g space where pixels from the same plane instance have similar embeddings. Then,
 the plane instances are obtained by grouping the embedding vectors in planar re
gions via an efficient mean shift clustering algorithm. In the second stage, we
estimate the parameter for each plane instance by considering both pixel-level a
nd instance-level consistencies. With the proposed method, we are able to detect
 an arbitrary number of planes. Extensive experiments on public datasets validat
e the effectiveness and efficiency of our method. Furthermore, our method runs a
t 30 fps at the testing time, thus could facilitate many real-time applications
such as visual SLAM and human-robot interaction. Code is available at https://gi
thub.com/svip-lab/PlanarReconstruction.
****

3DN: 3D Deformation Network
Weiyue Wang, Duygu Ceylan, Radomir Mech, Ulrich Neumann; Proceedings of the I
EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp.
1038-1046
Applications in virtual and augmented reality create a demand for rapid creation
 and easy access to large sets of 3D models. An effective way to address this de
mand is to edit or deform existing 3D models based on a reference, e.g., a 2D im
age which is very easy to acquire. Given such a source 3D model and a target whi
ch can be a 2D image, 3D model, or a point cloud acquired as a depth scan, we in
troduce 3DN, an end-to-end network that deforms the source model to resemble the
 target. Our method infers per-vertex offset displacements while keeping the mes
h connectivity of the source model fixed. We present a training strategy which u
ses a novel differentiable operation, mesh sampling operator, to generalize our
method across source and target models with varying mesh densities. Mesh samplin
g operator can be seamlessly integrated into the network to handle meshes with d
ifferent topologies. Qualitative and quantitative results show that our method g
enerates higher quality results compared to the state-of-the art learning-based
methods for 3D shape generation.
****

HorizonNet: Learning Room Layout With 1D Representation and Pano Stretch Data Au
gmentation
Cheng Sun, Chi-Wei Hsiao, Min Sun, Hwann-Tzong Chen; Proceedings of the IEEE/

CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1047-1056

We present a new approach to the problem of estimating the 3D room layout from a single panoramic image. We represent room layout as three 1D vectors that encode, at each image column, the boundary positions of floor-wall and ceiling-wall, and the existence of wall-wall boundary. The proposed network, HorizonNet, trained for predicting 1D layout, outperforms previous state-of-the-art approaches. The designed post-processing procedure for recovering 3D room layouts from 1D predictions can automatically infer the room shape with low computation cost--it takes less than 20ms for a panorama image while prior works might need dozens of seconds. We also propose Pano Stretch Data Augmentation, which can diversify panorama data and be applied to other panorama-related learning tasks. Due to the limited data available for non-cuboid layout, we relabel 65 general layout from the current dataset for finetuning. Our approach shows good performance on general layouts by qualitative results and cross-validation.

****

Deep Fitting Degree Scoring Network for Monocular 3D Object Detection

Lijie Liu, Jiwen Lu, Chunjing Xu, Qi Tian, Jie Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1057-1066

In this paper, we propose to learn a deep fitting degree scoring network for monocular 3D object detection, which aims to score fitting degree between proposals and object conclusively. Different from most existing monocular frameworks which use tight constraint to get 3D location, our approach achieves high-precision localization through measuring the visual fitting degree between the projected 3D proposals and the object. We first regress the dimension and orientation of the object using an anchor-based method so that a suitable 3D proposal can be constructed. We propose FQNet, which can infer the 3D IoU between the 3D proposals and the object solely based on 2D cues. Therefore, during the detection process, we sample a large number of candidates in the 3D space and project these 3D bounding boxes on 2D image individually. The best candidate can be picked out by simply exploring the spatial overlap between proposals and the object, in the form of the output 3D IoU score of FQNet. Experiments on the KITTI dataset demonstrate the effectiveness of our framework.

****

Pushing the Envelope for RGB-Based Dense 3D Hand Pose Estimation via Neural Rendering

Seungryul Baek, Kwang In Kim, Tae-Kyun Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1067-1076

Estimating 3D hand meshes from single RGB images is challenging, due to intrinsic 2D-3D mapping ambiguities and limited training data. We adopt a compact parametric 3D hand model that represents deformable and articulated hand meshes. To achieve the model fitting to RGB images, we investigate and contribute in three ways: 1) Neural rendering: inspired by recent work on human body, our hand mesh estimator (HME) is implemented by a neural network and a differentiable renderer, supervised by 2D segmentation masks and 3D skeletons. HME demonstrates good performance for estimating diverse hand shapes and improves pose estimation accuracies. 2) Iterative testing refinement: Our fitting function is differentiable. We iteratively refine the initial estimate using the gradients, in the spirit of iterative model fitting methods like ICP. The idea is supported by the latest research on human body. 3) Self-data augmentation: collecting sized RGB-mesh (or segmentation mask)-skeleton triplets for training is a big hurdle. Once the model is successfully fitted to input RGB images, its meshes i.e. shapes and articulations, are realistic, and we augment view-points on top of estimated dense hand poses. Experiments using three RGB-based benchmarks show that our framework offers beyond state-of-the-art accuracy in 3D pose estimation, as well as recovers dense 3D hand shapes. Each technical component above meaningfully improves the accuracy in the ablation study.

****

Self-Supervised Learning of 3D Human Pose Using Multi-View Geometry

Muhammed Kocabas, Salih Karagoz, Emre Akbas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1077-1086

Training accurate 3D human pose estimators requires large amount of 3D ground-truth data which is costly to collect. Various weakly or self supervised pose estimation methods have been proposed due to lack of 3D data. Nevertheless, these methods, in addition to 2D ground-truth poses, require either additional supervision in various forms (e.g. unpaired 3D ground truth data, a small subset of labels) or the camera parameters in multiview settings. To address these problems, we present EpipolarPose, a self-supervised learning method for 3D human pose estimation, which does not need any 3D ground-truth data or camera extrinsics. During training, EpipolarPose estimates 2D poses from multi-view images, and then, utilizes epipolar geometry to obtain a 3D pose and camera geometry which are subsequently used to train a 3D pose estimator. We demonstrate the effectiveness of our approach on standard benchmark datasets (i.e. Human3.6M and MPI-INF-3DHP) where we set the new state-of-the-art among weakly/self-supervised methods. Furthermore, we propose a new performance measure Pose Structure Score (PSS) which is a scale invariant, structure aware measure to evaluate the structural plausibility of a pose with respect to its ground truth. Code and pretrained models are available at https://github.com/mkocabas/EpipolarPose

****

FSA-Net: Learning Fine-Grained Structure Aggregation for Head Pose Estimation From a Single Image

Tsun-Yi Yang, Yi-Ting Chen, Yen-Yu Lin, Yung-Yu Chuang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1087-1096

This paper proposes a method for head pose estimation from a single image. Previous methods often predict head poses through landmark or depth estimation and would require more computation than necessary. Our method is based on regression and feature aggregation. For having a compact model, we employ the soft stagewise regression scheme. Existing feature aggregation methods treat inputs as a bag of features and thus ignore their spatial relationship in a feature map. We propose to learn a fine-grained structure mapping for spatially grouping features before aggregation. The fine-grained structure provides part-based information and pooled values. By utilizing learnable and non-learnable importance over the spatial location, different model variants can be generated and form a complementary ensemble. Experiments show that our method outperforms the state-of-the-art methods including both the landmark-free ones and the ones based on landmark or depth estimation. With only a single RGB frame as input, our method even outperforms methods utilizing multi-modality information (RGB-D, RGB-Time) on estimating the yaw angle. Furthermore, the memory overhead of our model is 100 times smaller than those of previous methods.

****

Dense 3D Face Decoding Over 2500FPS: Joint Texture & Shape Convolutional Mesh Decoders

Yuxiang Zhou, Jiankang Deng, Irene Kotsia, Stefanos Zafeiriou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1097-1106

3D Morphable Models (3DMMs) are statistical models that represent facial texture and shape variations using a set of linear bases and more particular Principal Component Analysis (PCA). 3DMMs were used as statistical priors for reconstructing 3D faces from images by solving non-linear least square optimization problems. Recently, 3DMMs were used as generative models for training non-linear mappings (i.e., regressors) from image to the parameters of the models via Deep Convolutional Neural Networks (DCNNs). Nevertheless, all of the above methods use either fully connected layers or 2D convolutions on parametric unwrapped UV spaces leading to large networks with many parameters. In this paper, we present the first, to the best of our knowledge, non-linear 3DMMs by learning joint texture and shape auto-encoders using direct mesh convolutions. We demonstrate how these auto-encoders can be used to train very light-weight models that perform Coloured Mesh Decoding (CMD) in-the-wild at a speed of over 2500 FPS.

****
Does Learning Specific Features for Related Parts Help Human Pose Estimation?
Wei Tang, Ying Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1107-1116
Human pose estimation (HPE) is inherently a homogeneous multi-task learning problem, with the localization of each body part as a different task. Recent HPE approaches universally learn a shared representation for all parts, from which their locations are linearly regressed. However, our statistical analysis indicates not all parts are related to each other. As a result, such a sharing mechanism can lead to negative transfer and deteriorate the performance. This potential issue drives us to raise an interesting question. Can we identify related parts and learn specific features for them to improve pose estimation? Since unrelated tasks no longer share a high-level representation, we expect to avoid the adverse effect of negative transfer. In addition, more explicit structural knowledge, e.g., ankles and knees are highly related, is incorporated into the model, which helps resolve ambiguities in HPE. To answer this question, we first propose a data-driven approach to group related parts based on how much information they share. Then a part-based branching network (PBN) is introduced to learn representations specific to each part group. We further present a multi-stage version of this network to repeatedly refine intermediate features and pose estimates. Ablation experiments indicate learning specific features significantly improves the localization of occluded parts and thus benefits HPE. Our approach also outperforms all state-of-the-art methods on two benchmark datasets, with an outstanding advantage when occlusion occurs.
****
Linkage Based Face Clustering via Graph Convolution Network
Zhongdao Wang, Liang Zheng, Yali Li, Shengjin Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1117-1125
In this paper, we present an accurate and scalable approach to the face clustering task. We aim at grouping a set of faces by their potential identities. We formulate this task as a link prediction problem: a link exists between two faces if they are of the same identity. The key idea is that we find the local context in the feature space around an instance (face) contains rich information about the linkage relationship between this instance and its neighbors. By constructing sub-graphs around each instance as input data, which depict the local context, we utilize the graph convolution network (GCN) to perform reasoning and infer the likelihood of linkage between pairs in the sub-graphs. Experiments show that our method is more robust to the complex distribution of faces than conventional methods, yielding favorably comparable results to state-of-the-art methods on standard face clustering benchmarks, and is scalable to large datasets. Furthermore, we show that the proposed method does not need the number of clusters as prior, is aware of noises and outliers, and can be extended to a multi-view version for more accurate clustering accuracy.
****
Towards High-Fidelity Nonlinear 3D Face Morphable Model
Luan Tran, Feng Liu, Xiaoming Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1126-1135
Embedding 3D morphable basis functions into deep neural networks opens great potential for models with better representation power. However, to faithfully learn those models from an image collection, it requires strong regularization to overcome ambiguities involved in the learning process. This critically prevents us from learning high fidelity face models which are needed to represent face images in high level of details. To address this problem, this paper presents a novel approach to learn additional proxies as means to side-step strong regularizations, as well as, leverages to promote detailed shape/albedo. To ease the learning, we also propose to use a dual-pathway network, a carefully-designed architecture that brings a balance between global and local-based models. By improving the nonlinear 3D morphable model in both learning objective and network architecture, we present a model which is superior in capturing higher level of details tha

n the linear or its precedent nonlinear counterparts. As a result, our model ach ieves state-of-the-art performance on 3D face reconstruction by solely optimizin g latent representations.
****

RegularFace: Deep Face Recognition via Exclusive Regularization
Kai Zhao, Jingyi Xu, Ming-Ming Cheng; Proceedings of the IEEE/CVF Conference o n Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1136-1144
We consider the face recognition task where facial images of the same identity (person) is expected to be closer in the representation space, while different id entities be far apart. Several recent studies encourage the intra-class compactn ess by developing loss functions that penalize the variance of representations o f the same identity. In this paper, we propose the `exclusive regularization' t hat focuses on the other aspect of discriminability -- the inter-class separabi lity, which is neglected in many recent approaches. The proposed method, named RegularFace, explicitly distances identities by penalizing the angle between an identity and its nearest neighbor, resulting in discriminative face representati ons. Our method has intuitive geometric interpretation and presents unique benef its that are absent in previous works. Quantitative comparisons against prior me thods on several open benchmarks demonstrate the superiority of our method. In a ddition, our method is easy to implement and requires only a few lines of python code on modern deep learning frameworks.
****

BridgeNet: A Continuity-Aware Probabilistic Network for Age Estimation
Wanhua Li, Jiwen Lu, Jianjiang Feng, Chunjing Xu, Jie Zhou, Qi Tian; Procee dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP R), 2019, pp. 1145-1154
Age estimation is an important yet very challenging problem in computer vision. Existing methods for age estimation usually apply a divide-and-conquer strategy to deal with heterogeneous data caused by the non-stationary aging process. Howe ver, the facial aging process is also a continuous process, and the continuity r elationship between different components has not been effectively exploited. In this paper, we propose BridgeNet for age estimation, which aims to mine the cont inuous relation between age labels effectively. The proposed BridgeNet consists of local regressors and gating networks. Local regressors partition the data spa ce into multiple overlapping subspaces to tackle heterogeneous data and gating n etworks learn continuity aware weights for the results of local regressors by em ploying the proposed bridge-tree structure, which introduces bridge connections into tree models to enforce the similarity between neighbor nodes. Moreover, the se two components of BridgeNet can be jointly learned in an end-to-end way. We s how experimental results on the MORPH II, FG-NET and Chalearn LAP 2015 datasets and find that BridgeNet outperforms the state-of-the-art methods.
****

GANFIT: Generative Adversarial Network Fitting for High Fidelity 3D Face Reconst ruction
Baris Gecer, Stylianos Ploumpis, Irene Kotsia, Stefanos Zafeiriou; Proceeding s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1155-1164
In the past few years, a lot of work has been done towards reconstructing the 3D facial structure from single images by capitalizing on the power of Deep Convol utional Neural Networks (DCNNs). In the most recent works, differentiable render ers were employed in order to learn the relationship between the facial identity features and the parameters of a 3D morphable model for shape and texture. The texture features either correspond to components of a linear texture space or ar e learned by auto-encoders directly from in-the-wild images. In all cases, the q uality of the facial texture reconstruction of the state-of-the-art methods is s till not capable of modeling textures in high fidelity. In this paper, we take a radically different approach and harness the power of Generative Adversarial Ne tworks (GANs) and DCNNs in order to reconstruct the facial texture and shape fro m single images. That is, we utilize GANs to train a very powerful generator of facial texture in UV space. Then, we revisit the original 3D Morphable Models (3

DMMs) fitting approaches making use of non-linear optimization to find the optim al latent parameters that best reconstruct the test image but under a new perspe ctive. We optimize the parameters with the supervision of pretrained deep identi ty features through our end-to-end differentiable framework. We demonstrate exce llent results in photorealistic and identity preserving 3D face reconstructions and achieve for the first time, to the best of our knowledge, facial texture rec onstruction with high-frequency details.
****

Improving the Performance of Unimodal Dynamic Hand-Gesture Recognition With Mult imodal Training
Mahdi Abavisani, Hamid Reza Vaezi Joze, Vishal M. Patel; Proceedings of the IE EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1 165-1174
We present an efficient approach for leveraging the knowledge from multiple moda lities in training unimodal 3D convolutional neural networks (3D-CNNs) for the t ask of dynamic hand gesture recognition. Instead of explicitly combining multi modal information, which is commonplace in many state-of-the-art methods, we pro pose a different framework in which we embed the knowledge of multiple modalitie s in individual networks so that each unimodal network can achieve an improved p erformance. In particular, we dedicate separate networks per available modality and enforce them to collaborate and learn to develop networks with common seman tics and better representations. We introduce a "spatiotemporal semantic alignm ent" loss (SSA) to align the content of the features from different networks. I n addition, we regularize this loss with our proposed "focal regularization para meter" to avoid negative knowledge transfer. Experimental results show that ou r framework improves the test time recognition accuracy of unimodal networks, an d provides the state-of-the-art performance on various dynamic hand gesture reco gnition datasets.
****

Learning to Reconstruct People in Clothing From a Single RGB Camera
Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, Ge rard Pons-Moll; Proceedings of the IEEE/CVF Conference on Computer Vision and Pa ttern Recognition (CVPR), 2019, pp. 1175-1186
We present Octopus, a learning-based model to infer the personalized 3D shape of people from a few frames (1-8) of a monocular video in which the person is movi ng with a reconstruction accuracy of 4 to 5mm, while being orders of magnitude f aster than previous methods. From semantic segmentation images, our Octopus mode l reconstructs a 3D shape, including the parameters of SMPL plus clothing and ha ir in 10 seconds or less. The model achieves fast and accurate predictions base d on two key design choices. First, by predicting shape in a canonical T-pose sp ace, the network learns to encode the images of the person into pose-invariant l atent codes, where the information is fused. Second, based on the observation th at feed-forward predictions are fast but do not always align with the input imag es, we predict using both, bottom-up and top-down streams (one per view) allowin g information to flow in both directions. Learning relies only on synthetic 3D d ata. Once learned, Octopus can take a variable number of frames as input, and is able to reconstruct shapes even from a single image with an accuracy of 5mm. Re sults on 3 different datasets demonstrate the efficacy and accuracy of our appro ach.
****

Distilled Person Re-Identification: Towards a More Scalable System
Ancong Wu, Wei-Shi Zheng, Xiaowei Guo, Jian-Huang Lai; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11 87-1196
Person re-identification (Re-ID), for matching pedestrians across non-overlappin g camera views, has made great progress in supervised learning with abundant lab elled data. However, the scalability problem is the bottleneck for applications in large-scale systems. We consider the scalability problem of Re-ID from three aspects: (1) low labelling cost by reducing label amount, (2) low extension cost by reusing existing knowledge and (3) low testing computation cost by using lig

htweight models. The requirements render scalable Re-ID a challenging problem. To solve these problems in a unified system, we propose a Multi-teacher Adaptive Similarity Distillation Framework, which requires only a few labelled identities of target domain to transfer knowledge from multiple teacher models to a user-specified lightweight student model without accessing source domain data. We propose the Log-Euclidean Similarity Distillation Loss for Re-ID and further integrate the Adaptive Knowledge Aggregator to select effective teacher models to transfer target-adaptive knowledge. Extensive evaluations show that our method can extend with high scalability and the performance is comparable to the state-of-the-art unsupervised and semi-supervised Re-ID methods.

****

A Perceptual Prediction Framework for Self Supervised Event Segmentation
Sathyanarayanan N. Aakur, Sudeep Sarkar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1197-1206

Temporal segmentation of long videos is an important problem, that has largely been tackled through supervised learning, often requiring large amounts of annotated training data. In this paper, we tackle the problem of self-supervised temporal segmentation that alleviates the need for any supervision in the form of labels (full supervision) or temporal ordering (weak supervision). We introduce a self-supervised, predictive learning framework that draws inspiration from cognitive psychology to segment long, visually complex videos into constituent events. Learning involves only a single pass through the training data. We also introduce a new adaptive learning paradigm that helps reduce the effect of catastrophic forgetting in recurrent neural networks. Extensive experiments on three publicly available datasets - Breakfast Actions, 50 Salads, and INRIA Instructional Videos datasets show the efficacy of the proposed approach. We show that the proposed approach outperforms weakly-supervised and unsupervised baselines by up to 24% and achieves competitive segmentation results compared to fully supervised baselines with only a single pass through the training data. Finally, we show that the proposed self-supervised learning paradigm learns highly discriminating features to improve action recognition.

****

COIN: A Large-Scale Dataset for Comprehensive Instructional Video Analysis
Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, Jie Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1207-1216

There are substantial instruction videos on the Internet, which enables us to acquire knowledge for completing various tasks. However, most existing datasets for instruction video analysis have the limitations in diversity and scale, which makes them far from many real-world applications where more diverse activities occur. Moreover, it still remains a great challenge to organize and harness such data. To address these problems, we introduce a large-scale dataset called "COIN" for COmprehensive INstruction video analysis. Organized with a hierarchical structure, the COIN dataset contains 11,827 videos of 180 tasks in 12 domains (e.g., vehicles, gadgets, etc.) related to our daily life. With a new developed toolbox, all the videos are annotated effectively with a series of step descriptions and the corresponding temporal boundaries. Furthermore, we propose a simple yet effective method to capture the dependencies among different steps, which can be easily plugged into conventional proposal-based action detection methods for localizing important steps in instruction videos. In order to provide a benchmark for instruction video analysis, we evaluate plenty of approaches on the COIN dataset under different evaluation criteria. We expect the introduction of the COIN dataset will promote the future in-depth research on instruction video analysis for the community.

****

Recurrent Attentive Zooming for Joint Crowd Counting and Precise Localization
Chenchen Liu, Xinyu Weng, Yadong Mu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1217-1226

Crowd counting is a new frontier in computer vision with far-reaching applications particularly in social safety management. A majority of existing works adopt

a methodology that first estimates a person-density map and then calculates integral over this map to obtain the final count. As noticed by several prior investigations, the learned density map can significantly deviate from the true person density even though the final reported count is precise. This implies that the density map is unreliable for localizing crowd. To address this issue, this work proposes a novel framework that simultaneously solving two inherently related tasks - crowd counting and localization. The contributions are several-fold. First, our formulation is based on a crucial observation that localization tends to be inaccurate at high-density regions, and increasing the resolution is an effective albeit simple solution for improving localization. We thus propose Recurrent Attentive Zooming Network, which recurrently detects ambiguous image region and zooms it into high resolution for re-inspection. Second, the two tasks of counting and localization mutually reinforce each other. We propose an adaptive fusion scheme that effectively elevates the performance. Finally, a well-defined evaluation metric is proposed for the rarely-explored localization task. We conduct comprehensive evaluations on several crowd benchmarks, including the newly-developed large-scale UCF-QNRF dataset and demonstrate superior advantages over state-of-the-art methods.

****

An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-Based Action Recognition

Chenyang Si, Wentao Chen, Wei Wang, Liang Wang, Tieniu Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1227-1236

Skeleton-based action recognition is an important task that requires the adequate understanding of movement characteristics of a human action from the given skeleton sequence. Recent studies have shown that exploring spatial and temporal features of the skeleton sequence is vital for this task. Nevertheless, how to effectively extract discriminative spatial and temporal features is still a challenging problem. In this paper, we propose a novel Attention Enhanced Graph Convolutional LSTM Network (AGC-LSTM) for human action recognition from skeleton data. The proposed AGC-LSTM can not only capture discriminative features in spatial configuration and temporal dynamics but also explore the co-occurrence relationship between spatial and temporal domains. We also present a temporal hierarchical architecture to increase temporal receptive fields of the top AGC-LSTM layer, which boosts the ability to learn the high-level semantic representation and significantly reduces the computation cost. Furthermore, to select discriminative spatial information, the attention mechanism is employed to enhance information of key joints in each AGC-LSTM layer. Experimental results on two datasets are provided: NTU RGB+D dataset and Northwestern-UCLA dataset. The comparison results demonstrate the effectiveness of our approach and show that our approach outperforms the state-of-the-art methods on both datasets.

****

Graph Convolutional Label Noise Cleaner: Train a Plug-And-Play Action Classifier for Anomaly Detection

Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H. Li, Ge Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1237-1246

Video anomaly detection under weak labels is formulated as a typical multiple-instance learning problem in previous works. In this paper, we provide a new perspective, i.e., a supervised learning task under noisy labels. In such a viewpoint, as long as cleaning away label noise, we can directly apply fully supervised action classifiers to weakly supervised anomaly detection, and take maximum advantage of these well-developed classifiers. For this purpose, we devise a graph convolutional network to correct noisy labels. Based upon feature similarity and temporal consistency, our network propagates supervisory signals from high-confidence snippets to low-confidence ones. In this manner, the network is capable of providing cleaned supervision for action classifiers. During the test phase, we only need to obtain snippet-wise predictions from the action classifier without any extra post-processing. Extensive experiments on 3 datasets at different scal

es with 2 types of action classifiers demonstrate the efficacy of our method. Remarkably, we obtain the frame-level AUC score of 82.12% on UCF-Crime.
****

MAN: Moment Alignment Network for Natural Language Moment Retrieval via Iterative Graph Adjustment
Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, Larry S. Davis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1247-1257
This research strives for natural language moment retrieval in long, untrimmed video streams. The problem is not trivial especially when a video contains multiple moments of interests and the language describes complex temporal dependencies, which often happens in real scenarios. We identify two crucial challenges: semantic misalignment and structural misalignment. However, existing approaches treat different moments separately and do not explicitly model complex moment-wise temporal relations. In this paper, we present Moment Alignment Network (MAN), a novel framework that unifies the candidate moment encoding and temporal structural reasoning in a single-shot feed-forward network. MAN naturally assigns candidate moment representations aligned with language semantics over different temporal locations and scales. Most importantly, we propose to explicitly model moment-wise temporal relations as a structured graph and devise an iterative graph adjustment network to jointly learn the best structure in an end-to-end manner. We evaluate the proposed approach on two challenging public benchmarks DiDeMo and Charades-STA, where our MAN significantly outperforms the state-of-the-art by a large margin.
****

Less Is More: Learning Highlight Detection From Video Duration
Bo Xiong, Yannis Kalantidis, Deepti Ghadiyaram, Kristen Grauman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1258-1267
Highlight detection has the potential to significantly ease video browsing, but existing methods often suffer from expensive supervision requirements, where human viewers must manually identify highlights in training videos. We propose a scalable unsupervised solution that exploits video duration as an implicit supervision signal. Our key insight is that video segments from shorter user-generated videos are more likely to be highlights than those from longer videos, since users tend to be more selective about the content when capturing shorter videos. Leveraging this insight, we introduce a novel ranking framework that prefers segments from shorter videos, while properly accounting for the inherent noise in the (unlabeled) training data. We use it to train a highlight detector with 10M hashtagged Instagram videos. In experiments on two challenging public video highlight detection benchmarks, our method substantially improves the state-of-the-art for unsupervised highlight detection.
****

DMC-Net: Generating Discriminative Motion Cues for Fast Compressed Video Action Recognition
Zheng Shou, Xudong Lin, Yannis Kalantidis, Laura Sevilla-Lara, Marcus Rohrbach, Shih-Fu Chang, Zhicheng Yan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1268-1277
Motion has shown to be useful for video understanding, where motion is typically represented by optical flow. However, computing flow from video frames is very timeconsuming. Recent works directly leverage the motion vectors and residuals readily available in the compressed video to represent motion at no cost. While this avoids flow computation, it also hurts accuracy since the motion vector is noisy and has substantially reduced resolution, which makes it a less discriminative motion representation. To remedy these issues, we propose a lightweight generator network, which reduces noises in motion vectors and captures fine motion details, achieving a more Discriminative Motion Cue (DMC) representation. Since optical flow is a more accurate motion representation, we train the DMC generator to approximate flow using a reconstruction loss and a generative adversarial loss, jointly with the downstream action classification task. Extensive evaluation

s on three action recognition benchmarks (HMDB-51, UCF-101, and a subset of Kinetics) confirm the effectiveness of our method. Our full system, consisting of the generator and the classifier, is coined as DMC-Net which obtains high accuracy close to that of using flow and runs two orders of magnitude faster than using optical flow at inference time.

****

AdaFrame: Adaptive Frame Selection for Fast Video Recognition

Zuxuan Wu, Caiming Xiong, Chih-Yao Ma, Richard Socher, Larry S. Davis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1278-1287

We present AdaFrame, a framework that adaptively selects relevant frames on a per-input basis for fast video recognition. AdaFrame contains a Long Short-Term Memory network augmented with a global memory that provides context information for searching which frames to use over time. Trained with policy gradient methods, AdaFrame generates a prediction, determines which frame to observe next, and computes the utility, i.e., expected future rewards, of seeing more frames at each time step. At testing time, AdaFrame exploits predicted utilities to achieve adaptive lookahead inference such that the overall computational costs are reduced without incurring a decrease in accuracy. Extensive experiments are conducted on two large-scale video benchmarks, FCVID and ActivityNet. AdaFrame matches the performance of using all frames with only 8.21 and 8.65 frames on FCVID and ActivityNet, respectively. We further qualitatively demonstrate learned frame usage can indicate the difficulty of making classification decisions; easier samples need fewer frames while harder ones require more, both at instance-level within the same class and at class-level among different categories.

****

Spatio-Temporal Video Re-Localization by Warp LSTM

Yang Feng, Lin Ma, Wei Liu, Jiebo Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1288-1297

The need for efficiently finding the video content a user wants is increasing because of the erupting of user-generated videos on the Web. Existing keyword-based or content-based video retrieval methods usually determine what occurs in a video but not when and where. In this paper, we make an answer to the question of when and where by formulating a new task, namely spatio-temporal video re-localization. Specifically, given a query video and a reference video, spatio-temporal video re-localization aims to localize tubelets in the reference video such that the tubelets semantically correspond to the query. To accurately localize the desired tubelets in the reference video, we propose a novel warp LSTM network, which propagates the spatio-temporal information for a long period and thereby captures the corresponding long-term dependencies. Another issue for spatio-temporal video re-localization is the lack of properly labeled video datasets. Therefore, we reorganize the videos in the AVA dataset to form a new dataset for spatio-temporal video re-localization research. Extensive experimental results show that the proposed model achieves superior performances over the designed baselines on the spatio-temporal video re-localization task.

****

Completeness Modeling and Context Separation for Weakly Supervised Temporal Action Localization

Daochang Liu, Tingting Jiang, Yizhou Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1298-1307

Temporal action localization is crucial for understanding untrimmed videos. In this work, we first identify two underexplored problems posed by the weak supervision for temporal action localization, namely action completeness modeling and action-context separation. Then by presenting a novel network architecture and its training strategy, the two problems are explicitly looked into. Specifically, to model the completeness of actions, we propose a multi-branch neural network in which branches are enforced to discover distinctive action parts. Complete actions can be therefore localized by fusing activations from different branches. And to separate action instances from their surrounding context, we generate hard negative data for training using the prior that motionless video clips are unli

kely to be actions. Experiments performed on datasets THUMOS'14 and ActivityNet show that our framework outperforms state-of-the-art methods. In particular, the average mAP on ActivityNet v1.2 is significantly improved from 18.0% to 22.4%. Our code will be released soon.
****

Unsupervised Deep Tracking
Ning Wang, Yibing Song, Chao Ma, Wengang Zhou, Wei Liu, Houqiang Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1308-1317
We propose an unsupervised visual tracking method in this paper. Different from existing approaches using extensive annotated data for supervised learning, our CNN model is trained on large-scale unlabeled videos in an unsupervised manner. Our motivation is that a robust tracker should be effective in both the forward and backward predictions (i.e., the tracker can forward localize the target object in successive frames and backtrace to its initial position in the first frame). We build our framework on a Siamese correlation filter network, which is trained using unlabeled raw videos. Meanwhile, we propose a multiple-frame validation method and a cost-sensitive loss to facilitate unsupervised learning. Without bells and whistles, the proposed unsupervised tracker achieves the baseline accuracy of fully supervised trackers, which require complete and accurate labels during training. Furthermore, unsupervised framework exhibits a potential in leveraging unlabeled or weakly labeled data to further improve the tracking accuracy.
****

Tracking by Animation: Unsupervised Learning of Multi-Object Attentive Trackers
Zhen He, Jian Li, Daxue Liu, Hangen He, David Barber; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1318-1327
Online Multi-Object Tracking (MOT) from videos is a challenging computer vision task which has been extensively studied for decades. Most of the existing MOT algorithms are based on the Tracking-by-Detection (TBD) paradigm combined with popular machine learning approaches which largely reduce the human effort to tune algorithm parameters. However, the commonly used supervised learning approaches require the labeled data (e.g., bounding boxes), which is expensive for videos. Also, the TBD framework is usually suboptimal since it is not end-to-end, i.e., it considers the task as detection and tracking, but not jointly. To achieve both label-free and end-to-end learning of MOT, we propose a Tracking-by-Animation framework, where a differentiable neural model first tracks objects from input frames and then animates these objects into reconstructed frames. Learning is then driven by the reconstruction error through backpropagation. We further propose a Reprioritized Attentive Tracking to improve the robustness of data association. Experiments conducted on both synthetic and real video datasets show the potential of the proposed model. Our project page is publicly available at: https://github.com/zhen-he/tracking-by-animation
****

Fast Online Object Tracking and Segmentation: A Unifying Approach
Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, Philip H.S. Torr; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1328-1338
In this paper we illustrate how to perform both visual object tracking and semi-supervised video object segmentation, in real-time, with a single simple approach. Our method, dubbed SiamMask, improves the offline training procedure of popular fully-convolutional Siamese approaches for object tracking by augmenting their loss with a binary segmentation task. Once trained, SiamMask solely relies on a single bounding box initialisation and operates online, producing class-agnostic object segmentation masks and rotated bounding boxes at 55 frames per second. Despite its simplicity, versatility and fast speed, our strategy allows us to establish a new state-of-the-art among real-time trackers on VOT-2018, while at the same time demonstrating competitive performance and the best speed for the semi-supervised video object segmentation task on DAVIS-2016 and DAVIS-2017.
****

Object Tracking by Reconstruction With View-Specific Discriminative Correlation Filters
Ugur Kart, Alan Lukezic, Matej Kristan, Joni-Kristian Kamarainen, Jiri Matas ; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1339-1348
Standard RGB-D trackers treat the target as a 2D structure, which makes modelling appearance changes related even to out-of-plane rotation challenging. This limitation is addressed by the proposed long-term RGB-D tracker called OTR - Object Tracking by Reconstruction. OTR performs online 3D target reconstruction to facilitate robust learning of a set of view-specific discriminative correlation filters (DCFs). The 3D reconstruction supports two performance- enhancing features: (i) generation of an accurate spatial support for constrained DCF learning from its 2D projection and (ii) point-cloud based estimation of 3D pose change for selection and storage of view-specific DCFs which robustly localize the target after out-of-view rotation or heavy occlusion. Extensive evaluation on the Princeton RGB-D tracking and STC Benchmarks shows OTR outperforms the state-of-the-art by a large margin.
****

SoPhie: An Attentive GAN for Predicting Paths Compliant to Social and Physical Constraints
Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, Hamid Rezatofighi, Silvio Savarese; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1349-1358
This paper addresses the problem of path prediction for multiple interacting agents in a scene, which is a crucial step for many autonomous platforms such as self-driving cars and social robots. We present SoPhie; an interpretable framework based on Generative Adversarial Network (GAN), which leverages two sources of information, the path history of all the agents in a scene, and the scene context information, using images of the scene. To predict a future path for an agent, both physical and social information must be leveraged. Previous work has not been successful to jointly model physical and social interactions. Our approach blends a social attention mechanism with physical attention that helps the model to learn where to look in a large scene and extract the most salient parts of the image relevant to the path. Whereas, the social attention component aggregates information across the different agent interactions and extracts the most important trajectory information from the surrounding neighbors. SoPhie also takes advantage of GAN to generates more realistic samples and to capture the uncertain nature of the future paths by modeling its distribution. All these mechanisms enable our approach to predict socially and physically plausible paths for the agents and to achieve state-of-the-art performance on several different trajectory forecasting benchmarks.
****

Leveraging Shape Completion for 3D Siamese Tracking
Silvio Giancola, Jesus Zarzar, Bernard Ghanem; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1359-1368
Point clouds are challenging to process due to their sparsity, therefore autonomous vehicles rely more on appearance attributes than pure geometric features. However, 3D LIDAR perception can provide crucial information for urban navigation in challenging light or weather conditions. In this paper, we investigate the versatility of Shape Completion for 3D Object Tracking in LIDAR point clouds. We design a Siamese tracker that encodes model and candidate shapes into a compact latent representation. We regularize the encoding by enforcing the latent representation to decode into an object model shape. We observe that 3D object tracking and 3D shape completion complement each other. Learning a more meaningful latent representation shows better discriminatory capabilities, leading to improved tracking performance. We test our method on the KITTI Tracking set using car 3D bounding boxes. Our model reaches a 76.94% Success rate and 81.38% Precision for 3D Object Tracking, with the shape completion regularization leading to an improvement of 3% in both metrics.
****

Target-Aware Deep Tracking
Xin Li, Chao Ma, Baoyuan Wu, Zhenyu He, Ming-Hsuan Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1369-1378
Existing deep trackers mainly use convolutional neural networks pre-trained for the generic object recognition task for representations. Despite demonstrated successes for numerous vision tasks, the contributions of using pre-trained deep features for visual tracking are not as significant as that for object recognition. The key issue is that in visual tracking the targets of interest can be arbitrary object class with arbitrary forms. As such, pre-trained deep features are less effective in modeling these targets of arbitrary forms for distinguishing them from the background. In this paper, we propose a novel scheme to learn target-aware features, which can better recognize the targets undergoing significant appearance variations than pre-trained deep features. To this end, we develop a regression loss and a ranking loss to guide the generation of target-active and scale-sensitive features. We identify the importance of each convolutional filter according to the back-propagated gradients and select the target-aware features based on activations for representing the targets. The target-aware features are integrated with a Siamese matching network for visual tracking. Extensive experimental results show that the proposed algorithm performs favorably against the state-of-the-art methods in terms of accuracy and speed.
****
Spatiotemporal CNN for Video Object Segmentation
Kai Xu, Longyin Wen, Guorong Li, Liefeng Bo, Qingming Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1379-1388
In this paper, we present a unified, end-to-end trainable spatiotemporal CNN model for VOS, which consists of two branches, i.e., the temporal coherence branch and the spatial segmentation branch. Specifically, the temporal coherence branch pretrained in an adversarial fashion from unlabeled video data, is designed to capture the dynamic appearance and motion cues of video sequences to guide object segmentation. The spatial segmentation branch focuses on segmenting objects accurately based on the learned appearance and motion cues. To obtain accurate segmentation results, we design a coarse-to-fine process to sequentially apply a designed attention module on multi-scale feature maps, and concatenate them to produce the final prediction. In this way, the spatial segmentation branch is enforced to gradually concentrate on object regions. These two branches are jointly fine-tuned on video segmentation sequences in an end-to-end manner. Several experiments are carried out on three challenging datasets (i.e., DAVIS-2016, DAVIS-2017 and Youtube-Object) to show that our method achieves favorable performance against the state-of-the-arts. Code is available at https://github.com/longyin880815/STCNN.
****
Towards Rich Feature Discovery With Class Activation Maps Augmentation for Person Re-Identification
Wenjie Yang, Houjing Huang, Zhang Zhang, Xiaotang Chen, Kaiqi Huang, Shu Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1389-1398
The fundamental challenge of small inter-person variation requires Person Re-Identification (Re-ID) models to capture sufficient fine-grained information. This paper proposes to discover diverse discriminative visual cues without extra assistance, e.g., pose estimation, human parsing. Specifically, a Class Activation Maps (CAM) augmentation model is proposed to expand the activation scope of baseline Re-ID model to explore rich visual cues, where the backbone network is extended by a series of ordered branches which share the same input but output complementary CAM. A novel Overlapped Activation Penalty is proposed to force the new branch to pay more attention to the image regions less activated by the old ones, such that spatial diverse visual features can be discovered. The proposed model achieves state-of-the-art results on three person Re-ID benchmarks. Moreover, a visualization approach termed ranking activation map (RAM) is proposed to expl

icitly interpret the ranking results in the test stage, which gives qualitative validations of the proposed method.
****

Wide-Context Semantic Image Extrapolation
Yi Wang, Xin Tao, Xiaoyong Shen, Jiaya Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1399-1408
This paper studies the fundamental problem of extrapolating visual context using deep generative models, i.e., extending image borders with plausible structure and details. This seemingly easy task actually faces many crucial technical challenges and has its unique properties. The two major issues are size expansion and one-side constraints. We propose a semantic regeneration network with several special contributions and use multiple spatial related losses to address these issues. Our results contain consistent structures and high-quality textures. Extensive experiments are conducted on various possible alternatives and related methods. We also explore the potential of our method for various interesting applications that can benefit research in a variety of fields.
****

End-To-End Time-Lapse Video Synthesis From a Single Outdoor Image
Seonghyeon Nam, Chongyang Ma, Menglei Chai, William Brendel, Ning Xu, Seon Joo Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1409-1418
Time-lapse videos usually contain visually appealing content but are often difficult and costly to create. In this paper, we present an end-to-end solution to synthesize a time-lapse video from a single outdoor image using deep neural networks. Our key idea is to train a conditional generative adversarial network based on existing datasets of time-lapse videos and image sequences. We propose a multi-frame joint conditional generation framework to effectively learn the correlation between the illumination change of an outdoor scene and the time of the day. We further present a multi-domain training scheme for robust training of our generative models from two datasets with different distributions and missing time stamp labels. Compared to alternative time-lapse video synthesis algorithms, our method uses the timestamp as the control variable and does not require a reference video to guide the synthesis of the final output. We conduct ablation studies to validate our algorithm and compare with state-of-the-art techniques both qualitatively and quantitatively.
****

GIF2Video: Color Dequantization and Temporal Interpolation of GIF Images
Yang Wang, Haibin Huang, Chuan Wang, Tong He, Jue Wang, Minh Hoai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1419-1428
Graphics Interchange Format (GIF) is a highly portable graphics format that is ubiquitous on the Internet. Despite their small sizes, GIF images often contain undesirable visual artifacts such as flat color regions, false contours, color shift, and dotted patterns. In this paper, we propose GIF2Video, the first learning-based method for enhancing the visual quality of GIFs in the wild. We focus on the challenging task of GIF restoration by recovering information lost in the three steps of GIF creation: frame sampling, color quantization, and color dithering. We first propose a novel CNN architecture for color dequantization. It is built upon a compositional architecture for multi-step color correction, with a comprehensive loss function designed to handle large quantization errors. We then adapt the SuperSlomo network for temporal interpolation of GIF frames. We introduce two large datasets, namely GIF-Faces and GIF-Moments, for both training and evaluation. Experimental results show that our method can significantly improve the visual quality of GIFs, and outperforms direct baseline and state-of-the-art approaches.
****

Mode Seeking Generative Adversarial Networks for Diverse Image Synthesis
Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, Ming-Hsuan Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1429-1437

Most conditional generation tasks expect diverse outputs given a single conditio nal context. However, conditional generative adversarial networks (cGANs) often focus on the prior conditional information and ignore the input noise vectors, w hich contribute to the output variations. Recent attempts to resolve the mode co llapse issue for cGANs are usually task-specific and computationally expensive. In this work, we propose a simple yet effective regularization term to address t he mode collapse issue for cGANs. The proposed method explicitly maximizes the r atio of the distance between generated images with respect to the corresponding latent codes, thus encouraging the generators to explore more minor modes during training. This mode seeking regularization term is readily applicable to variou s conditional generation tasks without imposing training overhead or modifying t he original network structures. We validate the proposed algorithm on three cond itional image synthesis tasks including categorical generation, image-to-image t ranslation, and text-to-image synthesis with different baseline models. Both qua litative and quantitative results demonstrate the effectiveness of the proposed regularization method for improving diversity without loss of quality.

****

Pluralistic Image Completion

Chuanxia Zheng, Tat-Jen Cham, Jianfei Cai; Proceedings of the IEEE/CVF Confere nce on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1438-1447

Most image completion methods produce only one result for each masked input, alt hough there may be many reasonable possibilities. In this paper, we present an a pproach for pluralistic image completion - the task of generating multiple and d iverse plausible solutions for image completion. A major challenge faced by lear ning-based approaches is that usually only one ground truth training instance pe r label. As such, sampling from conditional VAEs still leads to minimal diversit y. To overcome this, we propose a novel and probabilistically principled framewo rk with two parallel paths. One is a reconstructive path that utilizes the only one given ground truth to get prior distribution of missing parts and rebuild th e original image from this distribution. The other is a generative path for whic h the conditional prior is coupled to the distribution obtained in the reconstru ctive path. Both are supported by GANs. We also introduce a new short+long term attention layer that exploits distant relations among decoder and encoder featur es, improving appearance consistency. When tested on datasets with buildings (Pa ris), faces (CelebA-HQ), and natural images (ImageNet), our method not only gene rated higherquality completion results, but also with multiple and diverse plaus ible outputs.

****

Salient Object Detection With Pyramid Attention and Salient Edges

Wenguan Wang, Shuyang Zhao, Jianbing Shen, Steven C. H. Hoi, Ali Borji; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2019, pp. 1448-1457

This paper presents a new method for detecting salient objects in images using c onvolutional neural networks (CNNs). The proposed network, named PAGE-Net, offer s two key contributions. The first is the exploitation of an essential pyramid a ttention structure for salient object detection. This enables the network to con centrate more on salient regions while considering multi-scale saliency informat ion. Such a stacked attention design provides a powerful tool to efficiently imp rove the representation ability of the corresponding network layer with an enlar ged receptive field. The second contribution lies in the emphasis on the importa nce of salient edges. Salient edge information offers a strong cue to better seg ment salient objects and refine object boundaries. To this end, our model is equ ipped with a salient edge detection module, which is learned for precise salient boundary estimation. This encourages better edge-preserving salient object segm entation. Exhaustive experiments confirm that the proposed pyramid attention and salient edges are effective for salient object detection. We show that our deep saliency model outperforms state-of-the-art approaches for several benchmarks w ith a fast processing speed (25fps on one GPU).

****

Latent Filter Scaling for Multimodal Unsupervised Image-To-Image Translation

Yazeed Alharbi, Neil Smith, Peter Wonka; Proceedings of the IEEE/CVF Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1458-1466
In multimodal unsupervised image-to-image translation tasks, the goal is to tran slate an image from the source domain to many images in the target domain. We pr esent a simple method that produces higher quality images than current state-of- the-art while maintaining the same amount of multimodal diversity. Previous meth ods follow the unconditional approach of trying to map the latent code directly to a full-size image. This leads to complicated network architectures with sever al introduced hyperparameters to tune. By treating the latent code as a modifier of the convolutional filters, we produce multimodal output while maintaining th e traditional Generative Adversarial Network (GAN) loss and without additional h yperparameters. The only tuning required by our method controls the tradeoff bet ween variability and quality of generated images. Furthermore, we achieve disent anglement between source domain content and target domain style for free as a by -product of our formulation. We perform qualitative and quantitative experiments showing the advantages of our method compared with the state-of-the art on mult iple benchmark image-to-image translation datasets.
****

Attention-Aware Multi-Stroke Style Transfer
Yuan Yao, Jianqiang Ren, Xuansong Xie, Weidong Liu, Yong-Jin Liu, Jun Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti on (CVPR), 2019, pp. 1467-1475
Neural style transfer has drawn considerable attention from both academic and in dustrial field. Although visual effect and efficiency have been significantly im proved, existing methods are unable to coordinate spatial distribution of visual attention between the content image and stylized image, or render diverse level of detail via different brush strokes. In this paper, we tackle these limitatio ns by developing an attention-aware multi-stroke style transfer model. We first propose to assemble self-attention mechanism into a style-agnostic reconstructio n autoencoder framework, from which the attention map of a content image can be derived. By performing multi-scale style swap on content features and style feat ures, we produce multiple feature maps reflecting different stroke patterns. A f lexible fusion strategy is further presented to incorporate the salient characte ristics from the attention map, which allows integrating multiple stroke pattern s into different spatial regions of the output image harmoniously. We demonstrat e the effectiveness of our method, as well as generate comparable stylized image s with multiple stroke patterns against the state-of-the-art methods.
****

Feedback Adversarial Learning: Spatial Feedback for Improving Generative Adversa rial Networks
Minyoung Huh, Shao-Hua Sun, Ning Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1476-1485
We propose feedback adversarial learning (FAL) framework that can improve existi ng generative adversarial networks by leveraging spatial feedback from the discr iminator. We formulate the generation task as a recurrent framework, in which th e discriminator's feedback is integrated into the feedforward path of the genera tion process. Specifically, the generator conditions on the discriminator's spat ial output response, and its previous generation to improve generation quality o ver time - allowing the generator to attend and fix its previous mistakes. To ef fectively utilize the feedback, we propose an adaptive spatial transform layer, which learns to spatially modulate feature maps from its previous generation and the error signal from the discriminator. We demonstrate that one can easily ada pt FAL to existing adversarial learning frameworks on a wide range of tasks, inc luding image generation, image-to-image translation, and voxel generation.
****

Learning Pyramid-Context Encoder Network for High-Quality Image Inpainting
Yanhong Zeng, Jianlong Fu, Hongyang Chao, Baining Guo; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 14 86-1494
High-quality image inpainting requires filling missing regions in a damaged imag

e with plausible content. Existing works either fill the regions by copying high-resolution patches or generating semantically-coherent patches from region context, while neglecting the fact that both visual and semantic plausibility are highly-demanded. In this paper, we propose a Pyramid-context Encoder Network (denoted as PEN-Net) for image inpainting by deep generative models. The proposed PEN-Net is built upon a U-Net structure with three tailored components, ie., a pyramid-context encoder, a multi-scale decoder, and an adversarial training loss. First, we adopt a U-Net as backbone which can encode the context of an image from high-resolution pixels into high-level semantic features, and decode the features reversely. Second, we propose a pyramid-context encoder, which progressively learns region affinity by attention from a high-level semantic feature map, and transfers the learned attention to its adjacent high-resolution feature map. As the missing content can be filled by attention transfer from deep to shallow in a pyramid fashion, both visual and semantic coherence for image inpainting can be ensured. Third, we further propose a multi-scale decoder with deeply-supervised pyramid losses and an adversarial loss. Such a design not only results in fast convergence in training, but more realistic results in testing. Extensive experiments on a broad range of datasets shows the superior performance of the proposed network.

****

Example-Guided Style-Consistent Image Synthesis From Semantic Labeling

Miao Wang, Guo-Ye Yang, Ruilong Li, Run-Ze Liang, Song-Hai Zhang, Peter M. Hall, Shi-Min Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1495-1504

Example-guided image synthesis aims to synthesize an image from a semantic label map and an exemplary image indicating style. We use the term "style" in this problem to refer to implicit characteristics of images, for example: in portraits "style" includes gender, racial identity, age, hairstyle; in full body pictures it includes clothing; in street scenes it refers to weather and time of day and such like. A semantic label map in these cases indicates facial expression, full body pose, or scene segmentation. We propose a solution to the example-guided image synthesis problem using conditional generative adversarial networks with style consistency. Our key contributions are (i) a novel style consistency discriminator to determine whether a pair of images are consistent in style; (ii) an adaptive semantic consistency loss; and (iii) a training data sampling strategy, for synthesizing style-consistent results to the exemplar. We demonstrate the efficiency of our method on face, dance and street view synthesis tasks.

****

MirrorGAN: Learning Text-To-Image Generation by Redescription

Tingting Qiao, Jing Zhang, Duanqing Xu, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1505-1514

Generating an image from a given text description has two goals: visual realism and semantic consistency. Although significant progress has been made in generating high-quality and visually realistic images using generative adversarial networks, guaranteeing semantic consistency between the text description and visual content remains very challenging. In this paper, we address this problem by proposing a novel global-local attentive and semantic-preserving text-to-image-to-text framework called MirrorGAN. MirrorGAN exploits the idea of learning text-to-image generation by redescription and consists of three modules: a semantic text embedding module (STEM), a global-local collaborative attentive module for cascaded image generation (GLAM), and a semantic text regeneration and alignment module (STREAM). STEM generates word- and sentence-level embeddings. GLAM has a cascaded architecture for generating target images from coarse to fine scales, leveraging both local word attention and global sentence attention to progressively enhance the diversity and semantic consistency of the generated images. STREAM seeks to regenerate the text description from the generated image, which semantically aligns with the given text description. Thorough experiments on two public benchmark datasets demonstrate the superiority of MirrorGAN over other representative state-of-the-art methods.

****
Light Field Messaging With Deep Photographic Steganography
Eric Wengrowski, Kristin Dana; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1515-1524
We develop Light Field Messaging (LFM), a process of embedding, transmitting, and receiving hidden information in video that is displayed on a screen and captured by a handheld camera. The goal of the system is to minimize perceived visual artifacts of the message embedding, while simultaneously maximizing the accuracy of message recovery on the camera side. LFM requires photographic steganography for embedding messages that can be displayed and camera-captured. Unlike digital steganography, the embedding requirements are significantly more challenging due to the combined effect of the screen's radiometric emittance function, the camera's sensitivity function, and the camera-display relative geometry. We devise and train a network to jointly learn a deep embedding and recovery algorithm that requires no multi-frame synchronization. A key novel component is the camera display transfer function (CDTF) to model the camera-display pipeline. To learn this CDTF we introduce a dataset (Camera-Display 1M) of 1,000,000 camera-captured images collected from 25 camera-display pairs. The result of this work is a high-performance real-time LFM system using consumer-grade displays and smartphone cameras.
****
Im2Pencil: Controllable Pencil Illustration From Photographs
Yijun Li, Chen Fang, Aaron Hertzmann, Eli Shechtman, Ming-Hsuan Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1525-1534
We propose a high-quality photo-to-pencil translation method with fine-grained control over the drawing style. This is a challenging task due to multiple stroke types (e.g., outline and shading), structural complexity of pencil shading (e.g., hatching), and the lack of aligned training data pairs. To address these challenges, we develop a two-branch model that learns separate filters for generating sketchy outlines and tonal shading from a collection of pencil drawings. We create training data pairs by extracting clean outlines and tonal illustrations from original pencil drawings using image filtering techniques, and we manually label the drawing styles. In addition, our model creates different pencil styles (e.g., line sketchiness and shading style) in a user-controllable manner. Experimental results on different types of pencil drawings show that the proposed algorithm performs favorably against existing methods in terms of quality, diversity and user evaluations.
****
When Color Constancy Goes Wrong: Correcting Improperly White-Balanced Images
Mahmoud Afifi, Brian Price, Scott Cohen, Michael S. Brown; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1535-1544
This paper focuses on correcting a camera image that has been improperly white-balanced. This situation occurs when a camera's auto white balance fails or when the wrong manual white-balance setting is used. Even after decades of computational color constancy research, there are no effective solutions to this problem. The challenge lies not in identifying what the correct white balance should have been, but in the fact that the in-camera white-balance procedure is followed by several camera-specific nonlinear color manipulations that make it challenging to correct the image's colors in post-processing. This paper introduces the first method to explicitly address this problem. Our method is enabled by a dataset of over 65,000 pairs of incorrectly white-balanced images and their corresponding correctly white-balanced images. Using this dataset, we introduce a k-nearest neighbor strategy that is able to compute a nonlinear color mapping function to correct the image's colors. We show our method is highly effective and generalizes well to camera models not in the training set.
****
Beyond Volumetric Albedo -- A Surface Optimization Framework for Non-Line-Of-Sight Imaging

Chia-Yin Tsai, Aswin C. Sankaranarayanan, Ioannis Gkioulekas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1545-1555

Non-line-of-sight (NLOS) imaging is the problem of reconstructing properties of scenes occluded from a sensor, using measurements of light that indirectly travels from the occluded scene to the sensor through intermediate diffuse reflections. We introduce an analysis-by-synthesis framework that can reconstruct complex shape and reflectance of an NLOS object. Our framework deviates from prior work on NLOS reconstruction, by directly optimizing for a surface representation of the NLOS object, in place of commonly employed volumetric representations. At the core of our framework is a new rendering formulation that efficiently computes derivatives of radiometric measurements with respect to NLOS geometry and reflectance, while accurately modeling the underlying light transport physics. By coupling this with stochastic optimization and geometry processing techniques, we are able to reconstruct NLOS surface at a level of detail significantly exceeding what is possible with previous volumetric reconstruction methods.

****

Reflection Removal Using a Dual-Pixel Sensor

Abhijith Punnappurath, Michael S. Brown; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1556-1565

Reflection removal is the challenging problem of removing unwanted reflections that occur when imaging a scene that is behind a pane of glass. In this paper, we show that most cameras have an overlooked mechanism that can greatly simplify this task. Specifically, modern DLSR and smartphone cameras use dual pixel (DP) sensors that have two photodiodes per pixel to provide two sub-aperture views of the scene from a single captured image. "Defocus-disparity" cues, which are natural by-products of the DP sensor encoded within these two sub-aperture views, can be used to distinguish between image gradients belonging to the in-focus background and those caused by reflection interference. This gradient information can then be incorporated into an optimization framework to recover the background layer with higher accuracy than currently possible from the single captured image. As part of this work, we provide the first image dataset for reflection removal consisting of the sub-aperture views from the DP sensor.

****

Practical Coding Function Design for Time-Of-Flight Imaging

Felipe Gutierrez-Barragan, Syed Azer Reza, Andreas Velten, Mohit Gupta; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1566-1574

The depth resolution of a continuous-wave time-of-flight (CW-ToF) imaging system is determined by its coding functions. Recently, there has been growing interest in the design of new high-performance CW-ToF coding functions. However, these functions are typically designed in a hardware agnostic manner, i.e., without considering the practical device limitations, such as bandwidth, source power, digital (binary) function generation. Therefore, despite theoretical improvements, practical implementation of these functions remains a challenge. We present a constrained optimization approach for designing practical coding functions that adhere to hardware constraints. The optimization problem is non-convex with a large search space and no known globally optimal solutions. To make the problem tractable, we design an iterative, alternating least-squares algorithm, along with convex relaxation of the constraints. Using this approach, we design high-performance coding functions that can be implemented on existing hardware with minimal modifications. We demonstrate the performance benefits of the resulting functions via extensive simulations and a hardware prototype.

****

Meta-SR: A Magnification-Arbitrary Network for Super-Resolution

Xuecai Hu, Haoyuan Mu, Xiangyu Zhang, Zilei Wang, Tieniu Tan, Jian Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1575-1584

Recent research on super-resolution has achieved greatsuccess due to the development of deep convolutional neu-ral networks (DCNNs). However, super-resolution

of arbi-trary scale factor has been ignored for a long time. Mostprevious researchers regard super-resolution of differentscale factors as independent tasks. They train a specificmodel for each scale factor which is inefficient in comput-ing, and prior work only take the super-resolution of sev-eral integer scale factors into consideration. In this work,we propose a novel method called Meta-SR to firstly solvesuper-resolution of arbitrary scale factor (including non-integer scale factors) with a single model. In our Meta-SR,the Meta-Upscale Module is proposed to replace the tradi-tional upscale module. For arbitrary scale factor, the Meta-Upscale Module dynamically predicts the weights of the up-scale filters by taking the scale factor as input and use theseweights to generate the HR image of arbitrary size. For anylow-resolution image, our Meta-SR can continuously zoomin it with arbitrary scale factor by only using a single model.We evaluated the proposed method through extensive exper-iments on widely used benchmark datasets on single imagesuper-resolution. The experimental results show the superi-ority of our Meta-Upscale.

****

Multispectral and Hyperspectral Image Fusion by MS/HS Fusion Net

Qi Xie, Minghao Zhou, Qian Zhao, Deyu Meng, Wangmeng Zuo, Zongben Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1585-1594

Hyperspectral imaging can help better understand the characteristics of different materials, compared with traditional image systems. However, only high-resolution multispectral (HrMS) and low-resolution hyperspectral (LrHS) images can generally be captured at video rate in practice. In this paper, we propose a model-based deep learning approach for merging an HrMS and LrHS images to generate a high-resolution hyperspectral (HrHS) image. In specific, we construct a novel MS/HS fusion model which takes the observation models of low-resolution images and the low-rankness knowledge along the spectral mode of HrHS image into consideration. Then we design an iterative algorithm to solve the model by exploiting the proximal gradient method. And then, by unfolding the designed algorithm, we construct a deep network, called MS/HS Fusion Net, with learning the proximal operators and model parameters by convolutional neural networks. Experimental results on simulated and real data substantiate the superiority of our method both visually and quantitatively as compared with state-of-the-art methods along this line of research.

****

Learning Attraction Field Representation for Robust Line Segment Detection

Nan Xue, Song Bai, Fudong Wang, Gui-Song Xia, Tianfu Wu, Liangpei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1595-1603

This paper presents a region-partition based attraction field dual representation for line segment maps, and thus poses the problem of line segment detection (LSD) as the region coloring problem. The latter is then addressed by learning deep convolutional neural networks (ConvNets) for accuracy, robustness and efficiency. For a 2D line segment map, our dual representation consists of three components: (i) A region-partition map in which every pixel is assigned to one and only one line segment; (ii) An attraction field map in which every pixel in a partition region is encoded by its 2D projection vector w.r.t. the associated line segment; and (iii) A squeeze module which squashes the attraction field to a line segment map that almost perfectly recovers the input one. By leveraging the duality, we learn ConvNets to compute the attraction field maps for raw in-put images, followed by the squeeze module for LSD, in an end-to-end manner. Our method rigorously addresses several challenges in LSD such as local ambiguity and class imbalance. Our method also harnesses the best practices developed in ConvNets based semantic segmentation methods such as the encoder-decoder architecture and the a-trous convolution. In experiments, our method is tested on the WireFrame dataset and the YorkUrban dataset with state-of-the-art performance obtained. Especially, we advance the performance by 4.5 percents on the WireFramedataset. Our method is also fast with 6.6 10.4 FPS, outperforming most of existing line segment detectors.

****
Blind Super-Resolution With Iterative Kernel Correction
Jinjin Gu, Hannan Lu, Wangmeng Zuo, Chao Dong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1604-1613
Deep learning based methods have dominated super-resolution (SR) field due to their remarkable performance in terms of effectiveness and efficiency. Most of these methods assume that the blur kernel during downsampling is predefined/known (e.g., bicubic). However, the blur kernels involved in real applications are complicated and unknown, resulting in severe performance drop for the advanced SR methods. In this paper, we propose an Iterative Kernel Correction (IKC) method for blur kernel estimation in blind SR problem, where the blur kernels are unknown. We draw the observation that kernel mismatch could bring regular artifacts (either over-sharpening or over-smoothing), which can be applied to correct inaccurate blur kernels. Thus we introduce an iterative correction scheme -- IKC that achieves better results than direct kernel estimation. We further propose an effective SR network architecture using spatial feature transform (SFT) layers to handle multiple blur kernels, named SFTMD. Extensive experiments on synthetic and real-world images show that the proposed IKC method with SFTMD can provide visually favorable SR results and the state-of-the-art performance in blind SR problem.
****
Video Magnification in the Wild Using Fractional Anisotropy in Temporal Distribution
Shoichiro Takeda, Yasunori Akagi, Kazuki Okami, Megumi Isogai, Hideaki Kimata; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1614-1622
Video magnification methods can magnify and reveal subtle changes invisible to the naked eye. However, in such subtle changes, meaningful ones caused by physical and natural phenomena are mixed with non-meaningful ones caused by photographic noise. Therefore, current methods often produce noisy and misleading magnification outputs due to the non-meaningful subtle changes. For detecting only meaningful subtle changes, several methods have been proposed but require human manipulations, additional resources, or input video scene limitations. In this paper, we present a novel method using fractional anisotropy (FA) to detect only meaningful subtle changes without the aforementioned requirements. FA has been used in neuroscience to evaluate anisotropic diffusion of water molecules in the body. On the basis of our observation that temporal distribution of meaningful subtle changes more clearly indicates anisotropic diffusion than that of non-meaningful ones, we used FA to design a fractional anisotropic filter that passes only meaningful subtle changes. Using the filter enables our method to obtain better and more impressive magnification results than those obtained with state-of-the-art methods.
****
Attentive Feedback Network for Boundary-Aware Salient Object Detection
Mengyang Feng, Huchuan Lu, Errui Ding; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1623-1632
Recent deep learning based salient object detection methods achieve gratifying performance built upon Fully Convolutional Neural Networks (FCNs). However, most of them have suffered from the boundary challenge. The state-of-the-art methods employ feature aggregation tech- nique and can precisely find out wherein the salient object, but they often fail to segment out the entire object with fine boundaries, especially those raised narrow stripes. So there is still a large room for improvement over the FCN based models. In this paper, we design the Attentive Feedback Modules (AFMs) to better explore the structure of objects. A Boundary-Enhanced Loss (BEL) is further employed for learning exquisite boundaries. Our proposed deep model produces satisfying results on the object boundaries and achieves state-of-the-art performance on five widely tested salient object detection benchmarks. The network is in a fully convolutional fashion running at a speed of 26 FPS and does not need any post-processing.
****

Heavy Rain Image Restoration: Integrating Physics Model and Conditional Adversarial Learning
Ruoteng Li, Loong-Fah Cheong, Robby T. Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1633-1642
Most deraining works focus on rain streaks removal but they cannot deal adequately with heavy rain images. In heavy rain, streaks are strongly visible, dense rain accumulation or rain veiling effect significantly washes out the image, further scenes are relatively more blurry, etc. In this paper, we propose a novel method to address these problems. We put forth a 2-stage network: a physics-based backbone followed by a depth-guided GAN refinement. The first stage estimates the rain streaks, the transmission, and the atmospheric light governed by the underlying physics. To tease out these components more reliably, a guided filtering framework is used to decompose the image into its low- and high-frequency components. This filtering is guided by a rain-free residue image --- its content is used to set the passbands for the two channels in a spatially-variant manner so that the background details do not get mixed up with the rain-streaks. For the second stage, the refinement stage, we put forth a depth-guided GAN to recover the background details failed to be retrieved by the first stage, as well as correcting artefacts introduced by that stage. We have evaluated our method against state of the art methods. Extensive experiments show that our method outperforms them on real rain image data, recovering visually clean images with good details.
****
Learning to Calibrate Straight Lines for Fisheye Image Rectification
Zhucun Xue, Nan Xue, Gui-Song Xia, Weiming Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1643-1651
This paper presents a new deep-learning based method to simultaneously calibrate the intrinsic parameters of fisheye lens and rectify the distorted images. Assuming that the distorted lines generated by fisheye projection should be straight after rectification, we propose a novel deep neural network to impose explicit geometry constraints onto processes of the fisheye lens calibration and the distorted image rectification. In addition, considering the nonlinearity of distortion distribution in fisheye images, the proposed network fully exploits multi-scale perception to equalize the rectification effects on the whole image. To train and evaluate the proposed model, we also create a new large-scale dataset labeled with corresponding distortion parameters and well-annotated distorted lines. Compared with the state-of-the-art methods, our model achieves the best published rectification quality and the most accurate estimation of distortion parameters on a large set of synthetic and real fisheye images.
****
Camera Lens Super-Resolution
Chang Chen, Zhiwei Xiong, Xinmei Tian, Zheng-Jun Zha, Feng Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1652-1660
Existing methods for single image super-resolution (SR) are typically evaluated with synthetic degradation models such as bicubic or Gaussian downsampling. In this paper, we investigate SR from the perspective of camera lenses, named as CameraSR, which aims to alleviate the intrinsic tradeoff between resolution (R) and field-of-view (V) in realistic imaging systems. Specifically, we view the R-V degradation as a latent model in the SR process and learn to reverse it with realistic low- and high-resolution image pairs. To obtain the paired images, we propose two novel data acquisition strategies for two representative imaging systems (i.e., DSLR and smartphone cameras), respectively. Based on the obtained City100 dataset, we quantitatively analyze the performance of commonly-used synthetic degradation models, and demonstrate the superiority of CameraSR as a practical solution to boost the performance of existing SR methods. Moreover, CameraSR can be readily generalized to different content and devices, which serves as an advanced digital zoom tool in realistic imaging systems.
****

Frame-Consistent Recurrent Video Deraining With Dual-Level Flow
Wenhan Yang, Jiaying Liu, Jiashi Feng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1661-1670
In this paper, we address the problem of rain removal from videos by proposing a more comprehensive framework that considers the additional degradation factors in real scenes neglected in previous works. The proposed framework is built upon a two-stage recurrent network with dual-level flow regularizations to perform the inverse recovery process of the rain synthesis model for video deraining. The rain-free frame is estimated from the single rain frame at the first stage. It is then taken as guidance along with previously recovered clean frames to help obtain a more accurate clean frame at the second stage. This two-step architecture is capable of extracting more reliable motion information from the initially estimated rain-free frame at the first stage for better frame alignment and motion modeling at the second stage. Furthermore, to keep the motion consistency between frames that facilitates a frame-consistent deraining model at the second stage, a dual-level flow based regularization is proposed at both coarse flow and fine pixel levels. To better train and evaluate the proposed video deraining network, a novel rain synthesis model is developed to produce more visually authentic paired training and evaluation videos. Extensive experiments on a series of synthetic and real videos verify not only the superiority of the proposed method over state-of-the-art but also the effectiveness of network design and its each component.
****
Deep Plug-And-Play Super-Resolution for Arbitrary Blur Kernels
Kai Zhang, Wangmeng Zuo, Lei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1671-1681
While deep neural networks (DNN) based single image super-resolution (SISR) methods are rapidly gaining popularity, they are mainly designed for the widely-used bicubic degradation, and there still remains the fundamental challenge for them to super-resolve low-resolution (LR) image with arbitrary blur kernels. In the meanwhile, plug-and-play image restoration has been recognized with high flexibility due to its modular structure for easy plug-in of denoiser priors. In this paper, we propose a principled formulation and framework by extending bicubic degradation based deep SISR with the help of plug-and-play framework to handle LR images with arbitrary blur kernels. Specifically, we design a new SISR degradation model so as to take advantage of existing blind deblurring methods for blur kernel estimation. To optimize the new degradation induced energy function, we then derive a plug-and-play algorithm via variable splitting technique, which allows us to plug any super-resolver prior rather than the denoiser prior as a modular part. Quantitative and qualitative evaluations on synthetic and real LR images demonstrate that the proposed deep plug-and-play super-resolution framework is flexible and effective to deal with blurry LR images.
****
Sea-Thru: A Method for Removing Water From Underwater Images
Derya Akkaynak, Tali Treibitz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1682-1691
Robust recovery of lost colors in underwater images remains a challenging problem. We recently showed that this was partly due to the prevalent use of an atmospheric image formation model for underwater images. We proposed a physically accurate model that explicitly showed: 1) the attenuation coefficient of the signal is not uniform across the scene but depends on object range and reflectance, 2) the coefficient governing the increase in backscatter with distance differs from the signal attenuation coefficient. Here, we present a method that recovers color with the revised model using RGBD images. The Sea-thru method first calculates backscatter using the darkest pixels in the image and their known range information. Then, it uses an estimate of the spatially varying illuminant to obtain the range-dependent attenuation coefficient. Using more than 1,100 images from two optically different water bodies, which we make available, we show that our method outperforms those using the atmospheric model. Consistent removal of water will open up large underwater datasets to powerful computer vision and machine l

earning algorithms, creating exciting opportunities for the future of underwater exploration and conservation.
****

Deep Network Interpolation for Continuous Imagery Effect Transition
Xintao Wang, Ke Yu, Chao Dong, Xiaoou Tang, Chen Change Loy; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1692-1701
Deep convolutional neural network has demonstrated its capability of learning a deterministic mapping for the desired imagery effect. However, the large variety of user flavors motivates the possibility of continuous transition among different output effects. Unlike existing methods that require a specific design to achieve one particular transition (e.g., style transfer), we propose a simple yet universal approach to attain a smooth control of diverse imagery effects in many low-level vision tasks, including image restoration, image-to-image translation, and style transfer. Specifically, our method, namely Deep Network Interpolation (DNI), applies linear interpolation in the parameter space of two or more correlated networks. A smooth control of imagery effects can be achieved by tweaking the interpolation coefficients. In addition to DNI and its broad applications, we also investigate the mechanism of network interpolation from the perspective of learned filters.
****

Spatially Variant Linear Representation Models for Joint Filtering
Jinshan Pan, Jiangxin Dong, Jimmy S. Ren, Liang Lin, Jinhui Tang, Ming-Hsuan Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1702-1711
Joint filtering mainly uses an additional guidance image as a prior and transfers its structures to the target image in the filtering process. Different from existing algorithms that rely on locally linear models or hand-designed objective functions to extract the structural information from the guidance image, we propose a new joint filter based on a spatially variant linear representation model (SVLRM), where the target image is linearly represented by the guidance image. However, the SVLRM leads to a highly ill-posed problem. To estimate the linear representation coefficients, we develop an effective algorithm based on a deep convolutional neural network (CNN). The proposed deep CNN (constrained by the SVLRM) is able to estimate the spatially variant linear representation coefficients which are able to model the structural information of both the guidance and input images. We show that the proposed algorithm can be effectively applied to a variety of applications, including depth/RGB image upsampling and restoration, flash/no-flash image deblurring, natural image denoising, scale-aware filtering, etc. Extensive experimental results demonstrate that the proposed algorithm performs favorably against state-of-the-art methods that have been specially designed for each task.
****

Toward Convolutional Blind Denoising of Real Photographs
Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, Lei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1712-1722
While deep convolutional neural networks (CNNs) have achieved impressive success in image denoising with additive white Gaussian noise (AWGN), their performance remains limited on real-world noisy photographs. The main reason is that their learned models are easy to overfit on the simplified AWGN model which deviates severely from the complicated real-world noise model. In order to improve the generalization ability of deep CNN denoisers, we suggest training a convolutional blind denoising network (CBDNet) with more realistic noise model and real-world noisy-clean image pairs. On the one hand, both signal-dependent noise and in-camera signal processing pipeline is considered to synthesize realistic noisy images. On the other hand, real-world noisy photographs and their nearly noise-free counterparts are also included to train our CBDNet. To further provide an interactive strategy to rectify denoising result conveniently, a noise estimation subnetwork with asymmetric learning to suppress under-estimation of noise level is emb

edded into CBDNet. Extensive experimental results on three datasets of real-worl
d noisy pho- tographs clearly demonstrate the superior performance of CBDNet ove
r state-of-the-arts in terms of quantitative met- rics and visual quality. The c
ode has been made available at https://github.com/GuoShi28/CBDNet.
****

Towards Real Scene Super-Resolution With Raw Images
Xiangyu Xu, Yongrui Ma, Wenxiu Sun; Proceedings of the IEEE/CVF Conference on
Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1723-1731
Most existing super-resolution methods do not perform well in real scenarios due
to lack of realistic training data and information loss of the model input. To
solve the first problem, we propose a new pipeline to generate realistic trainin
g data by simulating the imaging process of digital cameras. And to remedy the i
nformation loss of the input, we develop a dual convolutional neural network to
exploit the originally captured radiance information in raw images. In addition,
we propose to learn a spatially-variant color transformation which helps more e
ffective color corrections. Extensive experiments demonstrate that super-resolut
ion with raw data helps recover fine details and clear structures, and more impo
rtantly, the proposed network and data generation pipeline achieve superior resu
lts for single image super-resolution in real scenarios.
****

ODE-Inspired Network Design for Single Image Super-Resolution
Xiangyu He, Zitao Mo, Peisong Wang, Yang Liu, Mingyuan Yang, Jian Cheng; Pr
oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
(CVPR), 2019, pp. 1732-1741
Single image super-resolution, as a high dimensional structured prediction probl
em, aims to characterize fine-grain information given a low-resolution sample. R
ecent advances in convolutional neural networks are introduced into super-resolu
tion and push forward progress in this field. Current studies have achieved impr
essive performance by manually designing deep residual neural networks but overl
y relies on practical experience. In this paper, we propose to adopt an ordinary
differential equation (ODE)-inspired design scheme for single image super-resol
ution, which have brought us a new understanding of ResNet in classification pro
blems. Not only is it interpretable for super-resolution but it provides a relia
ble guideline on network designs. By casting the numerical schemes in ODE as blu
eprints, we derive two types of network structures: LF-block and RK-block, which
correspond to the Leapfrog method and Runge-Kutta method in numerical ordinary
differential equations. We evaluate our models on benchmark datasets, and the re
sults show that our methods surpass the state-of-the-arts while keeping comparab
le parameters and operations.
****

Blind Image Deblurring With Local Maximum Gradient Prior
Liang Chen, Faming Fang, Tingting Wang, Guixu Zhang; Proceedings of the IEEE/
CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1742
-1750
Blind image deblurring aims to recover sharp image from a blurred one while the
blur kernel is unknown. To solve this ill-posed problem, a great amount of image
priors have been explored and employed in this area. In this paper, we present
a blind deblurring method based on Local Maximum Gradient (LMG) prior. Our work
is inspired by the simple and intuitive observation that the maximum value of a
local patch gradient will diminish after the blur process, which is proved to be
true both mathematically and empirically. This inherent property of blur proces
s helps us to establish a new energy function. By introducing an liner operator
to compute the Local Maximum Gradient, together with an effective optimization s
cheme, our method can handle various specific scenarios. Extensive experimental
results illustrate that our method is able to achieve favorable performance agai
nst state-of-the-art algorithms on both synthetic and real-world images.
****

Attention-Guided Network for Ghost-Free High Dynamic Range Imaging
Qingsen Yan, Dong Gong, Qinfeng Shi, Anton van den Hengel, Chunhua Shen, Ia
n Reid, Yanning Zhang; Proceedings of the IEEE/CVF Conference on Computer Visio

n and Pattern Recognition (CVPR), 2019, pp. 1751-1760
Ghosting artifacts caused by moving objects or misalignments is a key challenge in high dynamic range (HDR) imaging for dynamic scenes. Previous methods first register the input low dynamic range (LDR) images using optical flow before merging them, which are error-prone and cause ghosts in results. A very recent work tries to bypass optical flows via a deep network with skip-connections, however, which still suffers from ghosting artifacts for severe movement. To avoid the ghosting from the source, we propose a novel attention-guided end-to-end deep neural network (AHDRNet) to produce high-quality ghost-free HDR images. Unlike previous methods directly stacking the LDR images or features for merging, we use attention modules to guide the merging according to the reference image. The attention modules automatically suppress undesired components caused by misalignments and saturation and enhance desirable fine details in the non-reference images. In addition to the attention model, we use dilated residual dense block (DRDB) to make full use of the hierarchical features and increase the receptive field for hallucinating the missing details. The proposed AHDRNet is a non-flow-based method, which can also avoid the artifacts generated by optical-flow estimation error. Experiments on different datasets show that the proposed AHDRNet can achieve state-of-the-art quantitative and qualitative results.
****

Searching for a Robust Neural Architecture in Four GPU Hours
Xuanyi Dong, Yi Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1761-1770
Conventional neural architecture search (NAS) approaches are usually based on reinforcement learning or evolutionary strategy, which take more than 1000 GPU hours to find a good model on CIFAR-10. We propose an efficient NAS approach, which learns the searching approach by gradient descent. Our approach represents the search space as a directed acyclic graph (DAG). This DAG contains thousands of sub-graphs, each of which indicates a kind of neural architecture. To avoid traversing all the possibilities of the sub-graphs, we develop a differentiable sampler over the DAG. This sampler is learnable and optimized by the validation loss after training the sampled architecture. In this way, our approach can be trained in an end-to-end fashion by gradient descent, named Gradient-based search using Differentiable Architecture Sampler (GDAS). In experiments, we can finish one searching procedure in four GPU hours on CIFAR-10, and the discovered model obtains a test error of 2.82% with only 2.5M parameters, which is on par with the state-of-the-art.
****

Hierarchy Denoising Recursive Autoencoders for 3D Scene Layout Prediction
Yifei Shi, Angel X. Chang, Zhelun Wu, Manolis Savva, Kai Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1771-1780
Indoor scenes exhibit rich hierarchical structure in 3D object layouts. Many tasks in 3D scene understanding can benefit from reasoning jointly about the hierarchical context of a scene, and the identities of objects. We present a variational denoising recursive autoencoder (VDRAE) that generates and iteratively refines a hierarchical representation of 3D object layouts, interleaving bottom-up encoding for context aggregation and top-down decoding for propagation. We train our VDRAE on large-scale 3D scene datasets to predict both instance-level segmentations and a 3D object detections from an over-segmentation of an input point cloud. We show that our VDRAE improves object detection performance on real-world 3D point cloud datasets compared to baselines from prior work.
****

Adaptively Connected Neural Networks
Guangrun Wang, Keze Wang, Liang Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1781-1790
This paper presents a novel adaptively connected neural network (ACNet) to improve the traditional convolutional neural networks (CNNs) in two aspects. First, ACNet employs a flexible way to switch global and local inference in processing the internal feature representations by adaptively determining the connection s

tatus among the feature nodes (e.g., pixels of the feature maps). Note that in a computer vision domain, a node refers to a pixel of a feature map, while in the graph domain, a node denotes a graph node. We can show that existing CNNs, the classical multilayer perceptron (MLP), and the recently proposed non-local network (NLN) are all special cases of ACNet. Second, ACNet is also capable of handling non-Euclidean data. Extensive experimental analyses on a variety of benchmarks (i.e., ImageNet-1k classification, COCO 2017 detection and segmentation, CUHK03 person re-identification, CIFAR analysis, and Cora document categorization) demonstrate that ACNet cannot only achieve state-of-the-art performance but also overcome the limitation of the conventional MLP and CNN. The code is available at https://github.com/wanggrun/Adaptively-Connected-Neural-Networks.
****

CrDoCo: Pixel-Level Domain Transfer With Cross-Domain Consistency
Yun-Chun Chen, Yen-Yu Lin, Ming-Hsuan Yang, Jia-Bin Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1791-1800
Unsupervised domain adaptation algorithms aim to transfer the knowledge learned from one domain to another (e.g., synthetic to real images). The adapted representations often do not capture pixel-level domain shifts that are crucial for dense prediction tasks (e.g., semantic segmentation). In this paper, we present a novel pixel-wise adversarial domain adaptation algorithm. By leveraging image-to-image translation methods for data augmentation, our key insight is that while the translated images between domains may differ in styles, their predictions for the task should be consistent. We exploit this property and introduce a cross-domain consistency loss that enforces our adapted model to produce consistent predictions. Through extensive experimental results, we show that our method compares favorably against the state-of-the-art on a wide variety of unsupervised domain adaptation tasks.
****

Temporal Cycle-Consistency Learning
Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, Andrew Zisserman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1801-1810
We introduce a self-supervised representation learning method based on the task of temporal alignment between videos. The method trains a network using temporal cycle-consistency (TCC), a differentiable cycle-consistency loss that can be used to find correspondences across time in multiple videos. The resulting per-frame embeddings can be used to align videos by simply matching frames using nearest-neighbors in the learned embedding space. To evaluate the power of the embeddings, we densely label the Pouring and Penn Action video datasets for action phases. We show that (i) the learned embeddings enable few-shot classification of these action phases, significantly reducing the supervised training requirements; and (ii) TCC is complementary to other methods of self-supervised learning in videos, such as Shuffle and Learn and Time-Contrastive Networks. The embeddings are also used for a number of applications based on alignment (dense temporal correspondence) between video pairs, including transfer of metadata of synchronized modalities between videos (sounds, temporal semantic labels), synchronized playback of multiple videos, and anomaly detection. Project webpage: https://sites.google.com/view/temporal-cycle-consistency .
****

Predicting Future Frames Using Retrospective Cycle GAN
Yong-Hoon Kwon, Min-Gyu Park; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1811-1820
Recent advances in deep learning have significantly improved the performance of video prediction, however, top-performing algorithms start to generate blurry predictions as they attempt to predict farther future frames. In this paper, we propose a unified generative adversarial network for predicting accurate and temporally consistent future frames over time, even in a challenging environment. The key idea is to train a single generator that can predict both future and past frames while enforcing the consistency of bi-directional prediction using the ret

rospective cycle constraints. Moreover, we employ two discriminators not only to identify fake frames but also to distinguish fake contained image sequences from the real sequence. The latter discriminator, the sequence discriminator, plays a crucial role in predicting temporally consistent future frames. We experimentally verify the proposed framework using various real-world videos captured by car-mounted cameras, surveillance cameras, and arbitrary devices with state-of-the-art methods.

****

Density Map Regression Guided Detection Network for RGB-D Crowd Counting and Localization

Dongze Lian, Jing Li, Jia Zheng, Weixin Luo, Shenghua Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1821-1830

To simultaneously estimate head counts and localize heads with bounding boxes, a regression guided detection network (RDNet) is proposed for RGB-D crowd counting. Specifically, to improve the robustness of detection-based approaches for small/tiny heads, we leverage density map to improve the head/non-head classification in detection network where density map serves as the probability of a pixel being a head. A depth-adaptive kernel that considers the variances in head sizes is also introduced to generate high-fidelity density map for more robust density map regression. Further, a depth-aware anchor is designed for better initialization of anchor sizes in detection framework. Then we use the bounding boxes whose sizes are estimated with depth to train our RDNet. The existing RGB-D datasets are too small and not suitable for performance evaluation on data-driven based approaches, we collect a large-scale RGB-D crowd counting dataset. Experiments on both our RGB-D dataset and the MICC RGB-D counting dataset show that our method achieves the best performance for RGB-D crowd counting and localization. Further, our method can be readily extended to RGB image based crowd counting and achieves comparable performance on the ShanghaiTech Part_B dataset for both counting and localization.

****

TAFE-Net: Task-Aware Feature Embeddings for Low Shot Learning

Xin Wang, Fisher Yu, Ruth Wang, Trevor Darrell, Joseph E. Gonzalez; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1831-1840

Learning good feature embeddings for images often requires substantial training data. As a consequence, in settings where training data is limited (e.g., few-shot and zero-shot learning), we are typically forced to use a general feature embedding across prediction tasks. Ideally, we would like to construct feature embeddings that are tuned for the given task and even input image. In this work, we propose Task Aware Feature Embedding Networks (TAFE-Nets) to learn how to adapt the image representation to a new task in a meta learning fashion. Our network is composed of a meta learner and a prediction network, where the meta learner generates parameters for the feature layers in the prediction network based on a task input so that the feature embedding can be accurately adjusted for that task. We show that TAFE-Net is highly effective in generalizing to new tasks or concepts and evaluate the TAFE-Net on a range of benchmarks in zero-shot and few-shot learning. Our model matches or exceeds the state-of-the-art on all tasks. In particular, our approach improves the prediction accuracy of unseen attribute-object pairs by 4 to 15 points on the challenging visual attribute-object composition task.

****

Learning Semantic Segmentation From Synthetic Data: A Geometrically Guided Input-Output Adaptation Approach

Yuhua Chen, Wen Li, Xiaoran Chen, Luc Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1841-1850

As an alternative to manual pixel-wise annotation, synthetic data has been increasingly used for training semantic segmentation models. Such synthetic images and semantic labels can be easily generated from virtual 3D environments. In this work, we propose an approach to cross-domain semantic segmentation with the auxi

liary geometric information, which can also be easily obtained from virtual environments. The geometric information is utilized on two levels for reducing domain shift: on the input level, we augment the standard image translation network with the geometric information to translate synthetic images into realistic style; on the output level, we build a task network which simultaneously performs semantic segmentation and depth estimation. Meanwhile, adversarial training is applied on the joint output space to preserve the correlation between semantics and depth. The proposed approach is validated on two pairs of synthetic to real data set: from Virtual KITTI to KITTI, and from SYNTHIA to Cityscapes, where we achieve a clear performance gain compared to the baselines and various competing methods, demonstrating the effectiveness of the geometric information for cross-domain semantic segmentation.

****

Attentive Single-Tasking of Multiple Tasks

Kevis-Kokitsi Maninis, Ilija Radosavovic, Iasonas Kokkinos; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1851-1860

In this work we address task interference in universal networks by considering that a network is trained on multiple tasks, but performs one task at a time, an approach we refer to as "single-tasking multiple tasks". The network thus modifies its behaviour through task-dependent feature adaptation, or task attention. This gives the network the ability to accentuate the features that are adapted to a task, while shunning irrelevant ones. We further reduce task interference by forcing the task gradients to be statistically indistinguishable through adversarial training, ensuring that the common backbone architecture serving all tasks is not dominated by any of the task-specific gradients. Results in three multi-task dense labelling problems consistently show: (i) a large reduction in the number of parameters while preserving, or even improving performance and (ii) a smooth trade-off between computation and multi-task accuracy. We provide our system's code and pre-trained models at https://github.com/facebookresearch/astmt.

****

Deep Metric Learning to Rank

Fatih Cakir, Kun He, Xide Xia, Brian Kulis, Stan Sclaroff; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1861-1870

We propose a novel deep metric learning method by revisiting the learning to rank approach. Our method, named FastAP, optimizes the rank-based Average Precision measure, using an approximation derived from distance quantization. FastAP has a low complexity compared to existing methods, and is tailored for stochastic gradient descent. To fully exploit the benefits of the ranking formulation, we also propose a new minibatch sampling scheme, as well as a simple heuristic to enable large-batch training. On three few-shot image retrieval datasets, FastAP consistently outperforms competing methods, which often involve complex optimization heuristics or costly model ensembles.

****

End-To-End Multi-Task Learning With Attention

Shikun Liu, Edward Johns, Andrew J. Davison; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1871-1880

We propose a novel multi-task learning architecture, which allows learning of task-specific feature-level attention. Our design, the Multi-Task Attention Network (MTAN), consists of a single shared network containing a global feature pool, together with a soft-attention module for each task. These modules allow for learning of task-specific features from the global features, whilst simultaneously allowing for features to be shared across different tasks. The architecture can be trained end-to-end and can be built upon any feed-forward neural network, is simple to implement, and is parameter efficient. We evaluate our approach on a variety of datasets, across both image-to-image predictions and image classification tasks. We show that our architecture is state-of-the-art in multi-task learning compared to existing methods, and is also less sensitive to various weighting schemes in the multi-task loss function. Code is available at https://github.c

om/lorenmt/mtan.
****
Self-Supervised Learning via Conditional Motion Propagation
Xiaohang Zhan, Xingang Pan, Ziwei Liu, Dahua Lin, Chen Change Loy; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1881-1889
Intelligent agent naturally learns from motion. Various self-supervised algorithms have leveraged the motion cues to learn effective visual representations. The hurdle here is that motion is both ambiguous and complex, rendering previous works either suffer from degraded learning efficacy, or resort to strong assumptions on object motions. In this work, we design a new learning-from-motion paradigm to bridge these gaps. Instead of explicitly modeling the motion probabilities, we design the pretext task as a conditional motion propagation problem. Given an input image and several sparse flow guidance on it, our framework seeks to recover the full-image motion. Compared to other alternatives, our framework has several appealing properties: (1) Using sparse flow guidance during training resolves the inherent motion ambiguity, and thus easing feature learning. (2) Solving the pretext task of conditional motion propagation encourages the emergence of kinematically-sound representations that poss greater expressive power. Extensive experiments demonstrate that our framework learns structural and coherent features; and achieves state-of-the-art self-supervision performance on several downstream tasks including semantic segmentation, instance segmentation and human parsing. Furthermore, our framework is successfully extended to several useful applications such as semi-automatic pixel-level annotation.
****
Bridging Stereo Matching and Optical Flow via Spatiotemporal Correspondence
Hsueh-Ying Lai, Yi-Hsuan Tsai, Wei-Chen Chiu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1890-1899
Stereo matching and flow estimation are two essential tasks for scene understanding, spatially in 3D and temporally in motion. Existing approaches have been focused on the unsupervised setting due to the limited resource to obtain the large-scale ground truth data. To construct a self-learnable objective, co-related tasks are often linked together to form a joint framework. However, the prior work usually utilizes independent networks for each task, thus not allowing to learn shared feature representations across models. In this paper, we propose a single and principled network to jointly learn spatiotemporal correspondence for stereo matching and flow estimation, with a newly designed geometric connection as the unsupervised signal for temporally adjacent stereo pairs. We show that our method performs favorably against several state-of-the-art baselines for both unsupervised depth and flow estimation on the KITTI benchmark dataset.
****
All About Structure: Adapting Structural Information Across Domains for Boosting Semantic Segmentation
Wei-Lun Chang, Hui-Po Wang, Wen-Hsiao Peng, Wei-Chen Chiu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1900-1909
In this paper we tackle the problem of unsupervised domain adaptation for the task of semantic segmentation, where we attempt to transfer the knowledge learned upon synthetic datasets with ground-truth labels to real-world images without any annotation. With the hypothesis that the structural content of images is the most informative and decisive factor to semantic segmentation and can be readily shared across domains, we propose a Domain Invariant Structure Extraction (DISE) framework to disentangle images into domain-invariant structure and domain-specific texture representations, which can further realize image-translation across domains and enable label transfer to improve segmentation performance. Extensive experiments verify the effectiveness of our proposed DISE model and demonstrate its superiority over several state-of-the-art approaches.
****
Iterative Reorganization With Weak Spatial Constraints: Solving Arbitrary Jigsaw Puzzles for Unsupervised Representation Learning

Chen Wei, Lingxi Xie, Xutong Ren, Yingda Xia, Chi Su, Jiaying Liu, Qi Tian , Alan L. Yuille; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1910-1919

Learning visual features from unlabeled image data is an important yet challengi ng task, which is often achieved by training a model on some annotation-free inf ormation. We consider spatial contexts, for which we solve so-called jigsaw puzz les, i.e., each image is cut into grids and then disordered, and the goal is to recover the correct configuration. Existing approaches formulated it as a classi fication task by defining a fixed mapping from a small subset of configurations to a class set, but these approaches ignore the underlying relationship between different configurations and also limit their applications to more complex scena rios. This paper presents a novel approach which applies to jigsaw puzzles with an arbitrary grid size and dimensionality. We provide a fundamental and general ized principle, that weaker cues are easier to be learned in an unsupervised man ner and also transfer better. In the context of puzzle recognition, we use an it erative manner which, instead of solving the puzzle all at once, adjusts the ord er of the patches in each step until convergence. In each step, we combine both unary and binary features of each patch into a cost function judging the correct ness of the current configuration. Our approach, by taking similarity between pu zzles into consideration, enjoys a more efficient way of learning visual knowled ge. We verify the effectiveness of our approach from two aspects. First, it solv es arbitrarily complex puzzles, including high-dimensional puzzles, that prior m ethods are difficult to handle. Second, it serves as a reliable way of network i nitialization, which leads to better transfer performance in visual recognition tasks including classification, detection and segmentation.

****

Revisiting Self-Supervised Visual Representation Learning

Alexander Kolesnikov, Xiaohua Zhai, Lucas Beyer; Proceedings of the IEEE/CVF C onference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1920-1929

Unsupervised visual representation learning remains a largely unsolved problem i n computer vision research. Among a big body of recently proposed approaches for unsupervised learning of visual representations, a class of self-supervised tec hniques achieves superior performance on many challenging benchmarks. A large nu mber of the pretext tasks for self-supervised learning have been studied, but ot her important aspects, such as the choice of convolutional neural networks (CNN) , has not received equal attention. Therefore, we revisit numerous previously pr oposed self-supervised models, conduct a thorough large scale study and, as a re sult, uncover multiple crucial insights. We challenge a number of common practic es in self-supervised visual representation learning and observe that standard r ecipes for CNN design do not always translate to self-supervised representation learning. As part of our study, we drastically boost the performance of previous ly proposed techniques and outperform previously published state-of-the-art resu lts by a large margin. We will release the code for reproducing our experiments when the anonymity requirements are lifted.

****

It's Not About the Journey; It's About the Destination: Following Soft Paths Und er Question-Guidance for Visual Reasoning

Monica Haurilet, Alina Roitberg, Rainer Stiefelhagen; Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1930 -1939

Visual Reasoning remains a challenging task, as it has to deal with long-range a nd multi-step object relationships in the scene. We present a new model for Visu al Reasoning, aimed at capturing the interplay among individual objects in the i mage represented as a scene graph. As not all graph components are relevant for the query, we introduce the concept of a question-based visual guide, which cons trains the potential solution space by learning an optimal traversal scheme, whe re the final destination nodes alone are used to produce the answer. We show, th at finding relevant semantic structures facilitates generalization to new tasks by introducing a novel problem of knowledge transfer: training on one question t ype and answering questions from a different domain without any training data. F

urthermore, we report state-of-the-art results for Visual Reasoning on multiple query types and diverse image and video datasets.
****

Actively Seeking and Learning From Live Data
Damien Teney, Anton van den Hengel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1940-1949
One of the key limitations of traditional machine learning methods is their requirement for training data that exemplifies all the information to be learned. This is a particular problem for visual question answering methods, which may be asked questions about virtually anything. The approach we propose is a step toward overcoming this limitation by searching for the information required at test time. The resulting method dynamically utilizes data from an external source, such as a large set of questions/answers or images/captions. Concretely, we learn a set of base weights for a simple VQA model, that are specifically adapted to a given question with the information specifically retrieved for this question. The adaptation process leverages recent advances in gradient-based meta learning and contributions for efficient retrieval and cross-domain adaptation. We surpass the state-of-the-art on the VQA-CP v2 benchmark and demonstrate our approach to be intrinsically more robust to out-of-distribution test data. We demonstrate the use of external non-VQA data using the MS COCO captioning dataset to support the answering process. This approach opens a new avenue for open-domain VQA systems that interface with diverse sources of data.
****

Improving Referring Expression Grounding With Cross-Modal Attention-Guided Erasing
Xihui Liu, Zihao Wang, Jing Shao, Xiaogang Wang, Hongsheng Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1950-1959
Referring expression grounding aims at locating certain objects or persons in an image with a referring expression, where the key challenge is to comprehend and align various types of information from visual and textual domain, such as visual attributes, location and interactions with surrounding regions. Although the attention mechanism has been successfully applied for cross-modal alignments, previous attention models focus on only the most dominant features of both modalities, and neglect the fact that there could be multiple comprehensive textual-visual correspondences between images and referring expressions. To tackle this issue, we design a novel cross-modal attention-guided erasing approach, where we discard the most dominant information from either textual or visual domains to generate difficult training samples online, and to drive the model to discover complementary textual-visual correspondences. Extensive experiments demonstrate the effectiveness of our proposed method, which achieves state-of-the-art performance on three referring expression grounding datasets.
****

Neighbourhood Watch: Referring Expression Comprehension via Language-Guided Graph Attention Networks
Peng Wang, Qi Wu, Jiewei Cao, Chunhua Shen, Lianli Gao, Anton van den Hengel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1960-1968
The task in referring expression comprehension is to localize the object instance in an image described by a referring expression phrased in natural language. As a language-to-vision matching task, the key to this problem is to learn a discriminative object feature that can adapt to the expression used. To avoid ambiguity, the expression normally tends to describe not only the properties of the referent itself, but also its relationships to its neighbourhood. To capture and exploit this important information we propose a graph-based, language-guided attention mechanism. Being composed of node attention component and edge attention component, the proposed graph attention mechanism explicitly represents inter-object relationships, and properties with a flexibility and power impossible with competing approaches. Furthermore, the proposed graph attention mechanism enables the comprehension decision to be visualizable and explainable. Experiments on t

hree referring expression comprehension datasets show the advantage of the propo
sed approach.
****

Scene Graph Generation With External Knowledge and Image Reconstruction
Jiuxiang Gu, Handong Zhao, Zhe Lin, Sheng Li, Jianfei Cai, Mingyang Ling; P
roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
(CVPR), 2019, pp. 1969-1978
Scene graph generation has received growing attention with the advancements in i
mage understanding tasks such as object detection, attributes and relationship p
rediction, etc. However, existing datasets are biased in terms of object and rel
ationship labels, or often come with noisy and missing annotations, which makes
the development of a reliable scene graph prediction model very challenging. In
this paper, we propose a novel scene graph generation algorithm with external kn
owledge and image reconstruction loss to overcome these dataset issues. In parti
cular, we extract commonsense knowledge from the external knowledge base to refi
ne object and phrase features for improving generalizability in scene graph gene
ration. To address the bias of noisy object annotations, we introduce an auxilia
ry image reconstruction path to regularize the scene graph generation network. E
xtensive experiments show that our framework can generate better scene graphs, a
chieving the state-of-the-art performance on two benchmark datasets: Visual Rela
tionship Detection and Visual Genome datasets.
****

Polysemous Visual-Semantic Embedding for Cross-Modal Retrieval
Yale Song, Mohammad Soleymani; Proceedings of the IEEE/CVF Conference on Comput
er Vision and Pattern Recognition (CVPR), 2019, pp. 1979-1988
Visual-semantic embedding aims to find a shared latent space where related visua
l and textual instances are close to each other. Most current methods learn inje
ctive embedding functions that map an instance to a single point in the shared s
pace. Unfortunately, injective embedding cannot effectively handle polysemous in
stances with multiple possible meanings; at best, it would find an average repre
sentation of different meanings. This hinders its use in real-world scenarios wh
ere individual instances and their cross-modal associations are often ambiguous.
In this work, we introduce Polysemous Instance Embedding Networks (PIE-Nets) th
at compute multiple and diverse representations of an instance by combining glob
al context with locally-guided features via multi-head self-attention and residu
al learning. To learn visual-semantic embedding, we tie-up two PIE-Nets and opti
mize them jointly in the multiple instance learning framework. Most existing wor
k on cross-modal retrieval focus on image-text pairs of data. Here, we also tack
le a more challenging case of video-text retrieval. To facilitate further resear
ch in video-text retrieval, we release a new dataset of 50K video-sentence pairs
collected from social media, dubbed MRW (my reaction when). We demonstrate our
approach on both image-text and video-text retrieval scenarios using MS-COCO, TG
IF, and our new MRW dataset.
****

MUREL: Multimodal Relational Reasoning for Visual Question Answering
Remi Cadene, Hedi Ben-younes, Matthieu Cord, Nicolas Thome; Proceedings of th
e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, p
p. 1989-1998
Multimodal attentional networks are currently state-of-the-art models for Visual
Question Answering (VQA) tasks involving real images. Although attention allows
to focus on the visual content relevant to the question, this simple mechanism
is arguably insufficient to model complex reasoning features required for VQA or
other high-level tasks. In this paper, we propose MuRel, a multimodal relation
al network which is learned end-to-end to reason over real images. Our first con
tribution is the introduction of the MuRel cell, an atomic reasoning primitive r
epresenting interactions between question and image regions by a rich vectorial
representation, and modeling region relations with pairwise combinations. Second
ly, we incorporate the cell into a full MuRel network, which progressively refin
es visual and question interactions, and can be leveraged to define visualizatio
n schemes finer than mere attention maps. We validate the relevance of our appr

oach with various ablation studies, and show its superiority to attention-based methods on three datasets: VQA 2.0, VQA-CP v2 and TDIUC. Our final MuRel network is competitive to or outperforms state-of-the-art results in this challenging context. Our code is available: github.com/Cadene/murel.bootstrap.pytorch
****
Heterogeneous Memory Enhanced Multimodal Attention Model for Video Question Answering
Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, Heng Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1999-2007
In this paper, we propose a novel end-to-end trainable Video Question Answering (VideoQA) framework with three major components: 1) a new heterogeneous memory which can effectively learn global context information from appearance and motion features; 2) a redesigned question memory which helps understand the complex semantics of question and highlights queried subjects; and 3) a new multimodal fusion layer which performs multi-step reasoning by attending to relevant visual and textual hints with self-updated attention. Our VideoQA model firstly generates the global context-aware visual and textual features respectively by interacting current inputs with memory contents. After that, it makes the attentional fusion of the multimodal visual and textual representations to infer the correct answer. Multiple cycles of reasoning can be made to iteratively refine attention weights of the multimodal data and improve the final representation of the QA pair. Experimental results demonstrate our approach achieves state-of-the-art performance on four VideoQA benchmark datasets.
****
Information Maximizing Visual Question Generation
Ranjay Krishna, Michael Bernstein, Li Fei-Fei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2008-2018
Though image-to-sequence generation models have become overwhelmingly popular in human-computer communications, they suffer from strongly favoring safe generic questions ("What is in this picture?"). Generating uninformative but relevant questions is not sufficient or useful. We argue that a good question is one that has a tightly focused purpose --- one that is aimed at expecting a specific type of response. We build a model that maximizes mutual information between the image, the expected answer and the generated question. To overcome the non-differentiability of discrete natural language tokens, we introduce a variational continuous latent space onto which the expected answers project. We regularize this latent space with a second latent space that ensures clustering of similar answers. Even when we don't know the expected answer, this second latent space can generate goal-driven questions specifically aimed at extracting objects ("what is the person throwing"), attributes, ("What kind of shirt is the person wearing?"), color ("what color is the frisbee?"), material ("What material is the frisbee?"), etc. We quantitatively show that our model is able to retain information about an expected answer category, resulting in more diverse, goal-driven questions. We launch our model on a set of real world images and extract previously unseen visual concepts.
****
Learning to Detect Human-Object Interactions With Knowledge
Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, Mohan S. Kankanhalli; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2019-2028
The recent advances in instance-level detection tasks lay a strong foundation for automated visual scenes understanding. However, the ability to fully comprehend a social scene still eludes us. In this work, we focus on detecting human-object interactions (HOIs) in images, an essential step towards deeper scene understanding. HOI detection aims to localize human and objects, as well as to identify the complex interactions between them. Innate in practical problems with large label space, HOI categories exhibit a long-tail distribution, i.e., there exist some rare categories with very few training samples. Given the key observation that HOIs contain intrinsic semantic regularities despite they are visually diver

se, we tackle the challenge of long-tail HOI categories by modeling the underlyi ng regularities among verbs and objects in HOIs as well as general relationships . In particular, we construct a knowledge graph based on the ground-truth annota tions of training dataset and external source. In contrast to direct knowledge i ncorporation, we address the necessity of dynamic image-specific knowledge retri eval by multi-modal learning, which leads to an enhanced semantic embedding spac e for HOI comprehension. The proposed method shows improved performance on V-COC O and HICO-DET benchmarks, especially when predicting the rare HOI categories.
****

Learning Words by Drawing Images
Didac Suris, Adria Recasens, David Bau, David Harwath, James Glass, Antonio Torralba; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2029-2038
We propose a framework for learning through drawing. Our goal is to learn the co rrespondence between spoken words and abstract visual attributes, from a dataset of spoken descriptions of images. Building upon recent findings that GAN repres entations can be manipulated to edit semantic concepts in the generated output, we propose a new method to use such GAN-generated images to train a model using a triplet loss. To apply the method, we develop Audio CLEVRGAN, a new dataset o f audio descriptions of GAN-generated CLEVR images, and we describe a training p rocedure that creates a curriculum of GAN-generated images that focuses training on image pairs that differ in a specific, informative way. Training is done wit hout additional supervision beyond the spoken captions and the GAN. We find that training that takes advantage of GAN-generated edited examples results in impro vements in the model's ability to learn attributes compared to previous results. Our proposed learning framework also results in models that can associate spoke n words with some abstract visual concepts such as color and size.
****

Factor Graph Attention
Idan Schwartz, Seunghak Yu, Tamir Hazan, Alexander G. Schwing; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019 , pp. 2039-2048
Dialog is an effective way to exchange information, but subtle details and nuanc es are extremely important. While significant progress has paved a path to add ress visual dialog with algorithms, details and nuances remain a challenge. Att ention mechanisms have demonstrated compelling results to extract details in vis ual question answering and also provide a convincing framework for visual dialog due to their interpretability and effectiveness. However, the many data utiliti es that accompany visual dialog challenge existing attention techniques. We add ress this issue and develop a general attention mechanism for visual dialog whic h operates on any number of data utilities. To this end, we design a factor grap h based attention mechanism which combines any number of utility representations . We illustrate the applicability of the proposed approach on the challenging an d recently introduced VisDial datasets, outperforming recent state-of-the-art me thods by 1.1% for VisDial0.9 and by 2% for VisDial1.0 on MRR. Our ensemble model improved the MRR score on VisDial1.0 by more than 6%.
****

Reducing Uncertainty in Undersampled MRI Reconstruction With Active Acquisition
Zizhao Zhang, Adriana Romero, Matthew J. Muckley, Pascal Vincent, Lin Yang, Michal Drozdzal; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2049-2058
The goal of MRI reconstruction is to restore a high fidelity image from partiall y observed measurements. This partial view naturally induces reconstruction unce rtainty that can only be reduced by acquiring additional measurements. In this p aper, we present a novel method for MRI reconstruction that, at inference time, dynamically selects the measurements to take and iteratively refines the predict ion in order to best reduce the reconstruction error and, thus, its uncertainty. We validate our method on a large scale knee MRI dataset, as well as on ImageNe t. Results show that (1) our system successfully outperforms active acquisition baselines; (2) our uncertainty estimates correlate with error maps; and (3) our

ResNet-based architecture surpasses standard pixel-to-pixel models in the task of MRI reconstruction. The proposed method not only shows high-quality reconstructions but also paves the road towards more applicable solutions for accelerating MRI.

****

ESIR: End-To-End Scene Text Recognition via Iterative Image Rectification

Fangneng Zhan, Shijian Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2059-2068

Automated recognition of texts in scenes has been a research challenge for years, largely due to the arbitrary text appearance variation in perspective distortion, text line curvature, text styles and different types of imaging artifacts. The recent deep networks are capable of learning robust representations with respect to imaging artifacts and text style changes, but still face various problems while dealing with scene texts with perspective and curvature distortions. This paper presents an end-to-end trainable scene text recognition system (ESIR) that iteratively removes perspective distortion and text line curvature as driven by better scene text recognition performance. An innovative rectification network is developed, where a line-fitting transformation is designed to estimate the pose of text lines in scenes. Additionally, an iterative rectification framework is developed which corrects scene text distortions iteratively towards a fronto-parallel view. The ESIR is also robust to parameter initialization and easy to train, where the training needs only scene text images and word-level annotations as required by most scene text recognition systems. Extensive experiments over a number of public datasets show that the proposed ESIR is capable of rectifying scene text distortions accurately, achieving superior recognition performance for both normal scene text images and those suffering from perspective and curvature distortions.

****

ROI-10D: Monocular Lifting of 2D Detection to 6D Pose and Metric Shape

Fabian Manhardt, Wadim Kehl, Adrien Gaidon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2069-2078

We present a deep learning method for end-to-end monocular 3D object detection and metric shape retrieval. We propose a novel loss formulation by lifting 2D detection, orientation, and scale estimation into 3D space. Instead of optimizing these quantities separately, the 3D instantiation allows to properly measure the metric misalignment of boxes. We experimentally show that our 10D lifting of sparse 2D Regions of Interests (RoIs) achieves great results both for 6D pose and recovery of the textured metric geometry of instances. This further enables 3D synthetic data augmentation via inpainting recovered meshes directly onto the 2D scenes. We evaluate on KITTI3D against other strong monocular methods and demonstrate that our approach doubles the AP on the 3D pose metrics on the official test set, defining the new state of the art.

****

Collaborative Learning of Semi-Supervised Segmentation and Classification for Medical Images

Yi Zhou, Xiaodong He, Lei Huang, Li Liu, Fan Zhu, Shanshan Cui, Ling Shao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2079-2088

Medical image analysis has two important research areas: disease grading and fine-grained lesion segmentation. Although the former problem often relies on the latter, the two are usually studied separately. Disease severity grading can be treated as a classification problem, which only requires image-level annotations, while the lesion segmentation requires stronger pixel-level annotations. However, pixel-wise data annotation for medical images is highly time-consuming and requires domain experts. In this paper, we propose a collaborative learning method to jointly improve the performance of disease grading and lesion segmentation by semi-supervised learning with an attention mechanism. Given a small set of pixel-level annotated data, a multi-lesion mask generation model first performs the traditional semantic segmentation task. Then, based on initially predicted lesion maps for large quantities of image-level annotated data, a lesion attentive

disease grading model is designed to improve the severity classification accurac y. Meanwhile, the lesion attention model can refine the lesion maps using class-specific information to fine-tune the segmentation model in a semi-supervised ma nner. An adversarial architecture is also integrated for training. With extensiv e experiments on a representative medical problem called diabetic retinopathy (D R), we validate the effectiveness of our method and achieve consistent improveme nts over state-of-the-art methods on three public datasets.
****

Biologically-Constrained Graphs for Global Connectomics Reconstruction
Brian Matejek, Daniel Haehn, Haidong Zhu, Donglai Wei, Toufiq Parag, Hanspe ter Pfister; Proceedings of the IEEE/CVF Conference on Computer Vision and Patte rn Recognition (CVPR), 2019, pp. 2089-2098
Most current state-of-the-art connectome reconstruction pipelines have two major steps: initial pixel-based segmentation with affinity prediction and watershed transform, and refined segmentation by merging over-segmented regions. These met hods rely only on local context and are typically agnostic to the underlying bio logy. Since a few merge errors can lead to several incorrectly merged neuronal p rocesses, these algorithms are currently tuned towards over-segmentation produci ng an overburden of costly proofreading. We propose a third step for connectomic s reconstruction pipelines to refine an over-segmentation using both local and g lobal context with an emphasis on adhering to the underlying biology. We first e xtract a graph from an input segmentation where nodes correspond to segment labe ls and edges indicate potential split errors in the over-segmentation. In order to increase throughput and allow for large-scale reconstruction, we employ biolo gically inspired geometric constraints based on neuron morphology to reduce the number of nodes and edges. Next, two neural networks learn these neuronal shapes to further aid the graph construction process. Lastly, we reformulate the regio n merging problem as a graph partitioning one to leverage global context. We dem onstrate the performance of our approach on four real-world connectomics dataset s with an average variation of information improvement of 21.3%.
****

P3SGD: Patient Privacy Preserving SGD for Regularizing Deep CNNs in Pathological Image Classification
Bingzhe Wu, Shiwan Zhao, Guangyu Sun, Xiaolu Zhang, Zhong Su, Caihong Zeng, Zhihong Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat tern Recognition (CVPR), 2019, pp. 2099-2108
Recently, deep convolutional neural networks (CNNs) have achieved great success in pathological image classification. However, due to the limited number of labe led pathological images, there are still two challenges to be addressed: (1) ove rfitting: the performance of a CNN model is undermined by the overfitting due to its huge amounts of parameters and the insufficiency of labeled training data. (2) privacy leakage: the model trained using a conventional method may involunta rily reveal the private information of the patients in the training dataset. The smaller the dataset, the worse the privacy leakage. To tackle the above two ch allenges, we introduce a novel stochastic gradient descent (SGD) scheme, named p atient privacy preserving SGD (P3SGD), which performs the model update of the SG D in the patient level via a large-step update built upon each patient's data. S pecifically, to protect privacy and regularize the CNN model, we propose to inje ct the well-designed noise into the updates. Moreover, we equip our P3SGD with a n elaborated strategy to adaptively control the scale of the injected noise. To validate the effectiveness of P3SGD, we perform extensive experiments on a real-world clinical dataset and quantitatively demonstrate the superior ability of P3 SGD in reducing the risk of overfitting. We also provide a rigorous analysis of the privacy cost under differential privacy. Additionally, we find that the mode ls trained with P3SGD are resistant to the model-inversion attack compared with those trained using non-private SGD.
****

Elastic Boundary Projection for 3D Medical Image Segmentation
Tianwei Ni, Lingxi Xie, Huangjie Zheng, Elliot K. Fishman, Alan L. Yuille; P roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition

(CVPR), 2019, pp. 2109-2118
We focus on an important yet challenging problem: using a 2D deep network to deal with 3D segmentation for medical image analysis. Existing approaches either applied multi-view planar (2D) networks or directly used volumetric (3D) networks for this purpose, but both of them are not ideal: 2D networks cannot capture 3D contexts effectively, and 3D networks are both memory-consuming and less stable arguably due to the lack of pre-trained models. In this paper, we bridge the gap between 2D and 3D using a novel approach named Elastic Boundary Projection (EBP). The key observation is that, although the object is a 3D volume, what we really need in segmentation is to find its boundary which is a 2D surface. Therefore, we place a number of pivot points in the 3D space, and for each pivot, we determine its distance to the object boundary along a dense set of directions. This creates an elastic shell around each pivot which is initialized as a perfect sphere. We train a 2D deep network to determine whether each ending point falls within the object, and gradually adjust the shell so that it gradually converges to the actual shape of the boundary and thus achieves the goal of segmentation. EBP allows boundary-based segmentation without cutting a 3D volume into slices or patches, which stands out from conventional 2D and 3D approaches. EBP achieves promising accuracy in abdominal organ segmentation. Our code will be released on https://github.com/twni2016/Elastic-Boundary-Projection .
****

SIXray: A Large-Scale Security Inspection X-Ray Benchmark for Prohibited Item Discovery in Overlapping Images
Caijing Miao, Lingxi Xie, Fang Wan, Chi Su, Hongye Liu, Jianbin Jiao, Qixiang Ye; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2119-2128
In this paper, we present a large-scale dataset and establish a baseline for prohibited item discovery in Security Inspection X-ray images. Our dataset, named SIXray, consists of 1,059,231 X-ray images, in which 6 classes of 8,929 prohibited items are manually annotated. It raises a brand new challenge of overlapping image data, meanwhile shares the same properties with existing datasets, including complex yet meaningless contexts and class imbalance. We propose an approach named class-balanced hierarchical refinement (CHR) to deal with these difficulties. CHR assumes that each input image is sampled from a mixture distribution, and that deep networks require an iterative process to infer image contents accurately. To accelerate, we insert reversed connections to different network backbones, delivering high-level visual cues to assist mid-level features. In addition, a class-balanced loss function is designed to maximally alleviate the noise introduced by easy negative samples. We evaluate CHR on SIXray with different ratios of positive/negative samples. Compared to the baselines, CHR enjoys a better ability of discriminating objects especially using mid-level features, which offers the possibility of using a weakly-supervised approach towards accurate object localization. In particular, the advantage of CHR is more significant in the scenarios with fewer positive training samples, which demonstrates its potential application in real-world security inspection.
****

Noise2Void - Learning Denoising From Single Noisy Images
Alexander Krull, Tim-Oliver Buchholz, Florian Jug; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2129-2137
The field of image denoising is currently dominated by discriminative deep learning methods that are trained on pairs of noisy input and clean target images. Recently it has been shown that such methods can also be trained without clean targets. Instead, independent pairs of noisy images can be used, in an approach known as Noise2Noise (N2N). Here, we introduce Noise2Void (N2V), a training scheme that takes this idea one step further. It does not require noisy image pairs, nor clean target images. Consequently, N2V allows us to train directly on the body of data to be denoised and can therefore be applied when other methods cannot. Especially interesting is the application to biomedical image data, where the acquisition of training targets, clean or noisy, is frequently not possible. We co

mpare the performance of N2V to approaches that have either clean target images and/or noisy image pairs available. Intuitively, N2V cannot be expected to outperform methods that have more information available during training. Still, we observe that the denoising performance of Noise2Void drops in moderation and compares favorably to training-free denoising methods.
****

Joint Discriminative and Generative Learning for Person Re-Identification
Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, Jan Kautz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2138-2147
Person re-identification (re-id) remains challenging due to significant intra-class variations across different cameras. Recently, there has been a growing interest in using generative models to augment training data and enhance the invariance to input changes. The generative pipelines in existing methods, however, stay relatively separate from the discriminative re-id learning stages. Accordingly, re-id models are often trained in a straightforward manner on the generated data. In this paper, we seek to improve learned re-id embeddings by better leveraging the generated data. To this end, we propose a joint learning framework that couples re-id learning and data generation end-to-end. Our model involves a generative module that separately encodes each person into an appearance code and a structure code, and a discriminative module that shares the appearance encoder with the generative module. By switching the appearance or structure codes, the generative module is able to generate high-quality cross-id composed images, which are online fed back to the appearance encoder and used to improve the discriminative module. The proposed joint learning framework renders significant improvement over the baseline without using generated data, leading to the state-of-the-art performance on several benchmark datasets.
****

Unsupervised Person Re-Identification by Soft Multilabel Learning
Hong-Xing Yu, Wei-Shi Zheng, Ancong Wu, Xiaowei Guo, Shaogang Gong, Jian-Huang Lai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2148-2157
Although unsupervised person re-identification (RE-ID) has drawn increasing research attentions due to its potential to address the scalability problem of supervised RE-ID models, it is very challenging to learn discriminative information in the absence of pairwise labels across disjoint camera views. To overcome this problem, we propose a deep model for the soft multilabel learning for unsupervised RE-ID. The idea is to learn a soft multilabel (real-valued label likelihood vector) for each unlabeled person by comparing the unlabeled person with a set of known reference persons from an auxiliary domain. We propose the soft multilabel-guided hard negative mining to learn a discriminative embedding for the unlabeled target domain by exploring the similarity consistency of the visual features and the soft multilabels of unlabeled target pairs. Since most target pairs are cross-view pairs, we develop the cross-view consistent soft multilabel learning to achieve the learning goal that the soft multilabels are consistently good across different camera views. To enable effecient soft multilabel learning, we introduce the reference agent learning to represent each reference person by a reference agent in a joint embedding. We evaluate our unified deep model on Market-1501 and DukeMTMC-reID. Our model outperforms the state-of-the-art unsupervised RE-ID methods by clear margins. Code is available at https://github.com/KovenYu/MAR.
****

Learning Context Graph for Person Search
Yichao Yan, Qiang Zhang, Bingbing Ni, Wendong Zhang, Minghao Xu, Xiaokang Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2158-2167
Person re-identification has achieved great progress with deep convolutional neural networks. However, most previous methods focus on learning individual appearance feature embedding, and it is hard for the models to handle difficult situations with different illumination, large pose variance and occlusion. In this wor

k, we take a step further and consider employing context information for person search. For a probe-gallery pair, we first propose a contextual instance expansion module, which employs a relative attention module to search and filter useful context information in the scene. We also build a graph learning framework to effectively employ context pairs to update target similarity. These two modules are built on top of a joint detection and instance feature learning framework, which improves the discriminativeness of the learned features. The proposed framework achieves state-of-the-art performance on two widely used person search datasets.
****

Gradient Matching Generative Networks for Zero-Shot Learning
Mert Bulent Sariyildiz, Ramazan Gokberk Cinbis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2168-2178
Zero-shot learning (ZSL) is one of the most promising problems where substantial progress can potentially be achieved through unsupervised learning, due to distributional differences between supervised and zero-shot classes. For this reason, several works investigate the incorporation of discriminative domain adaptation techniques into ZSL, which, however, lead to modest improvements in ZSL accuracy. In contrast, we propose a generative model that can naturally learn from unsupervised examples, and synthesize training examples for unseen classes purely based on their class embeddings, and therefore, reduce the zero-shot learning problem into a supervised classification task. The proposed approach consists of two important components: (i) a conditional Generative Adversarial Network that learns to produce samples that mimic the characteristics of unsupervised data examples, and (ii) the Gradient Matching (GM) loss that measures the quality of the gradient signal obtained from the synthesized examples. Using our GM loss formulation, we enforce the generator to produce examples from which accurate classifiers can be trained. Experimental results on several ZSL benchmark datasets show that our approach leads to significant improvements over the state of the art in generalized zero-shot classification.
****

Doodle to Search: Practical Zero-Shot Sketch-Based Image Retrieval
Sounak Dey, Pau Riba, Anjan Dutta, Josep Llados, Yi-Zhe Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2179-2188
In this paper, we investigate the problem of zero-shot sketch-based image retrieval (ZS-SBIR), where human sketches are used as queries to conduct retrieval of photos from unseen categories. We importantly advance prior arts by proposing a novel ZS-SBIR scenario that represents a firm step forward in its practical application. The new setting uniquely recognizes two important yet often neglected challenges of practical ZS-SBIR, (i) the large domain gap between amateur sketch and photo, and (ii) the necessity for moving towards large-scale retrieval. We first contribute to the community a novel ZS-SBIR dataset, QuickDraw-Extended, that consists of 330,000 sketches and 204,000 photos spanning across 110 categories. Highly abstract amateur human sketches are purposefully sourced to maximize the domain gap, instead of ones included in existing datasets that can often be semi-photorealistic. We then formulate a ZS-SBIR framework to jointly model sketches and photos into a common embedding space. A novel strategy to mine the mutual information among domains is specifically engineered to alleviate the domain gap. External semantic knowledge is further embedded to aid semantic transfer. We show that, rather surprisingly, retrieval performance significantly outperforms that of state-of-the-art on existing datasets that can already be achieved using a reduced version of our model. We further demonstrate the superior performance of our full model by comparing with a number of alternatives on the newly proposed dataset. The new dataset, plus all training and testing code of our model, will be publicly released to facilitate future research.
****

Zero-Shot Task Transfer
Arghya Pal, Vineeth N Balasubramanian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2189-2198

In this work, we present a novel meta-learning algorithm that regresses model pa
rameters for novel tasks for which no ground truth is available (zero-shot tasks
). In order to adapt to novel zero-shot tasks, our meta-learner learns from the
model parameters of known tasks (with ground truth) and the correlation of known
 tasks to zero-shot tasks. Such intuition finds its foothold in cognitive scienc
e, where a subject (human baby) can adapt to a novel concept (depth understandin
g) by correlating it with old concepts (hand movement or self-motion), without r
eceiving an explicit supervision. We evaluated our model on the Taskonomy datase
t, with four tasks as zero-shot: surface normal, room layout, depth and camera p
ose estimation. These tasks were chosen based on the data acquisition complexity
 and the complexity associated with the learning process using a deep network. O
ur proposed methodolgy outperforms state-of-the-art models (which use ground tru
th) on each of our zero-shot tasks, showing promise on zero-shot task transfer.
We also conducted extensive experiments to study the various choices of our meth
odology, as well as showed how the proposed method can also be used in transfer
learning. To the best of our knowledge, this is the first such effort on zero-sh
ot learning in the task space.
****

C-MIL: Continuation Multiple Instance Learning for Weakly Supervised Object Dete
ction

Fang Wan, Chang Liu, Wei Ke, Xiangyang Ji, Jianbin Jiao, Qixiang Ye; Procee
dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP
R), 2019, pp. 2199-2208

Weakly supervised object detection (WSOD) is a challenging task when provided wi
th image category supervision but required to simultaneously learn object locati
ons and object detectors. Many WSOD approaches adopt multiple instance learning
(MIL) and have non-convex loss functions which are prone to get stuck into local
 minima (falsely localize object parts) while missing full object extent during
training. In this paper, we introduce a continuation optimization method into MI
L and thereby creating continuation multiple instance learning (C-MIL), with the
 intention of alleviating the non-convexity problem in a systematic way. We part
ition instances into spatially related and class related subsets, and approximat
e the original loss function with a series of smoothed loss functions defined wi
thin the subsets. Optimizing smoothed loss functions prevents the training proce
dure falling prematurely into local minima and facilitates the discovery of Stab
le Semantic Extremal Regions (SSERs) which indicate full object extent. On the P
ASCAL VOC 2007 and 2012 datasets, C-MIL improves the state-of-the-art of weakly
supervised object detection and weakly supervised object localization with large
 margins.
****

Weakly Supervised Learning of Instance Segmentation With Inter-Pixel Relations

Jiwoon Ahn, Sunghyun Cho, Suha Kwak; Proceedings of the IEEE/CVF Conference on
 Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2209-2218

This paper presents a novel approach for learning instance segmentation with ima
ge-level class labels as supervision. Our approach generates pseudo instance seg
mentation labels of training images, which are used to train a fully supervised
model. For generating the pseudo labels, we first identify confident seed areas
of object classes from attention maps of an image classification model, and prop
agate them to discover the entire instance areas with accurate boundaries. To th
is end, we propose IRNet, which estimates rough areas of individual instances an
d detects boundaries between different object classes. It thus enables to assign
 instance labels to the seeds and to propagate them within the boundaries so tha
t the entire areas of instances can be estimated accurately. Furthermore, IRNet
is trained with inter-pixel relations on the attention maps, thus no extra super
vision is required. Our method with IRNet achieves an outstanding performance on
 the PASCAL VOC 2012 dataset, surpassing not only previous state-of-the-art trai
ned with the same level of supervision, but also some of previous models relying
 on stronger supervision.
****

Attention-Based Dropout Layer for Weakly Supervised Object Localization

Junsuk Choe, Hyunjung Shim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2219-2228
Weakly Supervised Object Localization (WSOL) techniques learn the object location only using image-level labels, without location annotations. A common limitation for these techniques is that they cover only the most discriminative part of the object, not the entire object. To address this problem, we propose an Attention-based Dropout Layer (ADL), which utilizes the self-attention mechanism to process the feature maps of the model. The proposed method is composed of two key components: 1) hiding the most discriminative part from the model for capturing the integral extent of object, and 2) highlighting the informative region for improving the recognition power of the model. Based on extensive experiments, we demonstrate that the proposed method is effective to improve the accuracy of WSOL, achieving a new state-of-the-art localization accuracy in CUB-200-2011 dataset. We also show that the proposed method is much more efficient in terms of both parameter and computation overheads than existing techniques.
****

Domain Generalization by Solving Jigsaw Puzzles
Fabio M. Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, Tatiana Tommasi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2229-2238
Human adaptability relies crucially on the ability to learn and merge knowledge both from supervised and unsupervised learning: the parents point out few important concepts, but then the children fill in the gaps on their own. This is particularly effective, because supervised learning can never be exhaustive and thus learning autonomously allows to discover invariances and regularities that help to generalize. In this paper we propose to apply a similar approach to the task of object recognition across domains: our model learns the semantic labels in a supervised fashion, and broadens its understanding of the data by learning from self-supervised signals how to solve a jigsaw puzzle on the same images. This secondary task helps the network to learn the concepts of spatial correlation while acting as a regularizer for the classification task. Multiple experiments on the PACS, VLCS, Office-Home and digits datasets confirm our intuition and show that this simple method outperforms previous domain generalization and adaptation solutions. An ablation study further illustrates the inner workings of our approach.
****

Transferrable Prototypical Networks for Unsupervised Domain Adaptation
Yingwei Pan, Ting Yao, Yehao Li, Yu Wang, Chong-Wah Ngo, Tao Mei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2239-2247
In this paper, we introduce a new idea for unsupervised domain adaptation via a remold of Prototypical Networks, which learn an embedding space and perform classification via a remold of the distances to the prototype of each class. Specifically, we present Transferrable Prototypical Networks (TPN) for adaptation such that the prototypes for each class in source and target domains are close in the embedding space and the score distributions predicted by prototypes separately on source and target data are similar. Technically, TPN initially matches each target example to the nearest prototype in the source domain and assigns an example a "pseudo" label. The prototype of each class could then be computed on source-only, target-only and source-target data, respectively. The optimization of TPN is end-to-end trained by jointly minimizing the distance across the prototypes on three types of data and KL-divergence of score distributions output by each pair of the prototypes. Extensive experiments are conducted on the transfers across MNIST, USPS and SVHN datasets, and superior results are reported when comparing to state-of-the-art approaches. More remarkably, we obtain an accuracy of 80.4% of single model on VisDA 2017 dataset.
****

Blending-Target Domain Adaptation by Adversarial Meta-Adaptation Networks
Ziliang Chen, Jingyu Zhuang, Xiaodan Liang, Liang Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 22

48-2257
(Unsupervised) Domain Adaptation (DA) seeks for classifying target instances when solely provided with source labeled and target unlabeled examples for training. Learning domain-invariant features helps to achieve this goal, whereas it underpins unlabeled samples drawn from a single or multiple explicit target domains (Multi-target DA). In this paper, we consider a more realistic transfer scenario: our target domain is comprised of multiple sub-targets implicitly blended with each other so that learners could not identify which sub-target each unlabeled sample belongs to. This Blending-target Domain Adaptation (BTDA) scenario commonly appears in practice and threatens the validities of existing DA algorithms, due to the presence of domain gaps and categorical misalignments among these hidden sub-targets. To reap the transfer performance gains in this new scenario, we propose Adversarial Meta-Adaptation Network (AMEAN). AMEAN entails two adversarial transfer learning processes. The first is a conventional adversarial transfer to bridge our source and mixed target domains. To circumvent the intra-target category misalignment, the second process presents as "learning to adapt": It deploys an unsupervised meta-learner receiving target data and their ongoing feature-learning feedbacks, to discover target clusters as our "meta-sub-target" domains. This meta-sub-targets auto-design our meta-sub-target adaptation loss, which is capable to progressively eliminate the implicit category mismatching in our mixed target. We evaluate AMEAN and a variety of DA algorithms in three benchmarks under the BTDA setup. Empirical results show that BTDA is a quite challenging transfer setup for most existing DA algorithms, yet AMEAN significantly outperforms these state-of-the-art baselines and effectively restrains the negative transfer effects in BTDA.
****

ELASTIC: Improving CNNs With Dynamic Scaling Policies
Huiyu Wang, Aniruddha Kembhavi, Ali Farhadi, Alan L. Yuille, Mohammad Rastegari; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2258-2267
Scale variation has been a challenge from traditional to modern approaches in computer vision. Most solutions to scale issues have a similar theme: a set of intuitive and manually designed policies that are generic and fixed (e.g. SIFT or feature pyramid). We argue that the scaling policy should be learned from data. In this paper, we introduce Elastic, a simple, efficient and yet very effective approach to learn a dynamic scale policy from data. We formulate the scaling policy as a non-linear function inside the network's structure that (a) is learned from data, (b) is instance specific, (c) does not add extra computation, and (d) can be applied on any network architecture. We applied Elastic to several state-of-the-art network architectures and showed consistent improvement without extra (sometimes even lower) computation on ImageNet classification, MSCOCO multi-label classification, and PASCAL VOC semantic segmentation. Our results show major improvement for images with scale challenges. Our code is available here: https://github.com/allenai/elastic
****

ScratchDet: Training Single-Shot Object Detectors From Scratch
Rui Zhu, Shifeng Zhang, Xiaobo Wang, Longyin Wen, Hailin Shi, Liefeng Bo, Tao Mei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2268-2277
Current state-of-the-art object objectors are fine-tuned from the off-the-shelf networks pretrained on large-scale classification dataset ImageNet, which incurs some additional problems: 1) The classification and detection have different degrees of sensitivity to translation, resulting in the learning objective bias; 2) The architecture is limited by the classification network, leading to the inconvenience of modification. To cope with these problems, training detectors from scratch is a feasible solution. However, the detectors trained from scratch generally perform worse than the pretrained ones, even suffer from the convergence issue in training. In this paper, we explore to train object detectors from scratch robustly. By analysing the previous work on optimization landscape, we find that one of the overlooked points in current trained-from-scratch detector is the

BatchNorm. Resorting to the stable and predictable gradient brought by BatchNorm, detectors can be trained from scratch stably while keeping the favourable performance independent to the network architecture. Taking this advantage, we are able to explore various types of networks for object detection, without suffering from the poor convergence. By extensive experiments and analyses on downsampling factor, we propose the Root-ResNet backbone network, which makes full use of the information from original images. Our ScratchDet achieves the state-of-the-art accuracy on PASCAL VOC 2007, 2012 and MS COCO among all the train-from-scratch detectors and even performs better than several one-stage pretrained methods. Codes will be made publicly available at https://github.com/KimSoybean/ScratchDet.

****

SFNet: Learning Object-Aware Semantic Correspondence

Junghyup Lee, Dohyung Kim, Jean Ponce, Bumsub Ham; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2278-2287

We address the problem of semantic correspondence, that is, establishing a dense flow field between images depicting different instances of the same object or scene category. We propose to use images annotated with binary foreground masks and subjected to synthetic geometric deformations to train a convolutional neural network (CNN) for this task. Using these masks as part of the supervisory signal offers a good compromise between semantic flow methods, where the amount of training data is limited by the cost of manually selecting point correspondences, and semantic alignment ones, where the regression of a single global geometric transformation between images may be sensitive to image-specific details such as background clutter. We propose a new CNN architecture, dubbed SFNet, which implements this idea. It leverages a new and differentiable version of the argmax function for end-to-end training, with a loss that combines mask and flow consistency with smoothness terms. Experimental results demonstrate the effectiveness of our approach, which significantly outperforms the state of the art on standard benchmarks.

****

Deep Metric Learning Beyond Binary Supervision

Sungyeon Kim, Minkyo Seo, Ivan Laptev, Minsu Cho, Suha Kwak; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2288-2297

Metric Learning for visual similarity has mostly adopted binary supervision indicating whether a pair of images are of the same class or not. Such a binary indicator covers only a limited subset of image relations, and is not sufficient to represent semantic similarity between images described by continuous and/or structured labels such as object poses, image captions, and scene graphs. Motivated by this, we present a novel method for deep metric learning using continuous labels. First, we propose a new triplet loss that allows distance ratios in the label space to be preserved in the learned metric space. The proposed loss thus enables our model to learn the degree of similarity rather than just the order. Furthermore, we design a triplet mining strategy adapted to metric learning with continuous labels. We address three different image retrieval tasks with continuous labels in terms of human poses, room layouts and image captions, and demonstrate the superior performance of our approach compared to previous methods.

****

Learning to Cluster Faces on an Affinity Graph

Lei Yang, Xiaohang Zhan, Dapeng Chen, Junjie Yan, Chen Change Loy, Dahua Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2298-2306

Face recognition sees remarkable progress in recent years, and its performance has reached a very high level. Taking it to a next level requires substantially larger data, which would involve prohibitive annotation cost. Hence, exploiting unlabeled data becomes an appealing alternative. Recent works have shown that clustering unlabeled faces is a promising approach, often leading to notable performance gains. Yet, how to effectively cluster, especially on a large-scale (i.e.

million-level or above) dataset, remains an open question. A key challenge lies in the complex variations of cluster patterns, which make it difficult for conventional clustering methods to meet the needed accuracy. This work explores a novel approach, namely, learning to cluster instead of relying on hand-crafted criteria. Specifically, we propose a framework based on graph convolutional network, which combines a detection and a segmentation module to pinpoint face clusters. Experiments show that our method yields significantly more accurate face clusters, which, as a result, also lead to further performance gain in face recognition.

****

C2AE: Class Conditioned Auto-Encoder for Open-Set Recognition

Poojan Oza, Vishal M. Patel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2307-2316

Models trained for classification often assume that all testing classes are known while training. As a result, when presented with an unknown class during testing, such closed-set assumption forces the model to classify it as one of the known classes. However, in a real world scenario, classification models are likely to encounter such examples. Hence, identifying those examples as unknown becomes critical to model performance. A potential solution to overcome this problem lies in a class of learning problems known as open-set recognition. It refers to the problem of identifying the unknown classes during testing, while maintaining performance on the known classes. In this paper, we propose an open-set recognition algorithm using class conditioned auto-encoders with novel training and testing methodologies. In this method, training procedure is divided in two sub-tasks, 1. closed-set classification and, 2. open-set identification (i.e. identifying a class as known or unknown). Encoder learns the first task following the closed-set classification training pipeline, whereas decoder learns the second task by reconstructing conditioned on class identity. Furthermore, we model reconstruction errors using the Extreme Value Theory of statistical modeling to find the threshold for identifying known/unknown class samples. Experiments performed on multiple image classification datasets show that the proposed method performs significantly better than the state of the art methods. The source code is available at: github.com/otkupjnoz/c2ae.

****

Shapes and Context: In-The-Wild Image Synthesis & Manipulation

Aayush Bansal, Yaser Sheikh, Deva Ramanan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2317-2326

We introduce a data-driven model for interactively synthesizing in-the-wild images from semantic label input masks. Our approach is dramatically different from recent work in this space, in that we make use of no learning. Instead, our approach uses simple but classic tools for matching scene context, shapes, and parts to a stored library of exemplars. Though simple, this approach has several notable advantages over recent work: (1) because nothing is learned, it is not limited to specific training data distributions (such as cityscapes, facades, or faces); (2) it can synthesize arbitrarily high-resolution images, limited only by the resolution of the exemplar library; (3) by appropriately composing shapes and parts, it can generate an exponentially large set of viable candidate output images (that can say, be interactively searched by a user). We present results on the diverse COCO dataset, significantly outperforming learning-based approaches on standard image synthesis metrics. Finally, we explore user-interaction and user-controllability, demonstrating that our system can be used as a platform for user-driven content creation.

****

Semantics Disentangling for Text-To-Image Generation

Guojun Yin, Bin Liu, Lu Sheng, Nenghai Yu, Xiaogang Wang, Jing Shao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2327-2336

Synthesizing photo-realistic images from text descriptions is a challenging problem. Previous studies have shown remarkable progresses on visual quality of the generated images. In this paper, we consider semantics from the input text descr

iptions in helping render photo-realistic images. However, diverse linguistic ex
pressions pose challenges in extracting consistent semantics even they depict th
e same thing. To this end, we propose a novel photo-realistic text-to-image gene
ration model that implicitly disentangles semantics to both fulfill the high-lev
el semantic consistency and low-level semantic diversity. To be specific, we des
ign (1) a Siamese mechanism in the discriminator to learn consistent high-level
semantics, and (2) a visual-semantic embedding strategy by semantic-conditioned
batch normalization to find diverse low-level semantics. Extensive experiments a
nd ablation studies on CUB and MS-COCO datasets demonstrate the superiority of t
he proposed method in comparison to state-of-the-art methods.
****

Semantic Image Synthesis With Spatially-Adaptive Normalization
Taesung Park, Ming-Yu Liu, Ting-Chun Wang, Jun-Yan Zhu; Proceedings of the IE
EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2
337-2346
We propose spatially-adaptive normalization, a simple but effective layer for sy
nthesizing photorealistic images given an input semantic layout. Previous method
s directly feed the semantic layout as input to the network, forcing the network
to memorize the information throughout all the layers. Instead, we propose usin
g the input layout for modulating the activations in normalization layers throug
h a spatially-adaptive, learned affine transformation. Experiments on several c
hallenging datasets demonstrate the superiority of our method compared to existi
ng approaches, regarding both visual fidelity and alignment with input layouts.
Finally, our model allows users to easily control the style and content of image
synthesis results as well as create multi-modal results. Code is available upon
publication.
****

Progressive Pose Attention Transfer for Person Image Generation
Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, Xiang Bai; Pro
ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (
CVPR), 2019, pp. 2347-2356
This paper proposes a new generative adversarial network to the problem of pose
transfer, i.e., transferring the pose of a given person to a target one. The gen
erator of the network comprises a sequence of Pose-Attentional Transfer Blocks t
hat each transfers certain regions it attends to, generating the person image pr
ogressively. Compared with those in previous works, our generated person images
possess better appearance consistency and shape consistency with the input image
s, thus significantly more realistic-looking. The efficacy and efficiency of the
proposed network are validated both qualitatively and quantitatively on Market-
1501 and DeepFashion. Furthermore, the proposed architecture can generate traini
ng images for person re-identification, alleviating data insufficiency.
****

Unsupervised Person Image Generation With Semantic Parsing Transformation
Sijie Song, Wei Zhang, Jiaying Liu, Tao Mei; Proceedings of the IEEE/CVF Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2357-2366
In this paper, we address unsupervised pose-guided person image generation, whic
h is known challenging due to non-rigid deformation. Unlike previous methods lea
rning a rock-hard direct mapping between human bodies, we propose a new pathway
to decompose the hard mapping into two more accessible subtasks, namely, semanti
c parsing transformation and appearance generation. Firstly, a semantic generati
ve network is proposed to transform between semantic parsing maps, in order to s
implify the non-rigid deformation learning. Secondly, an appearance generative n
etwork learns to synthesize semantic-aware textures. Thirdly, we demonstrate tha
t training our framework in an end-to-end manner further refines the semantic ma
ps and final results accordingly. Our method is generalizable to other semantic-
aware person image generation tasks, e.g., clothing texture transfer and control
led image manipulation. Experimental results demonstrate the superiority of our
method on DeepFashion and Market-1501 datasets, especially in keeping the clothi
ng attributes and better body shapes.
****

DeepView: View Synthesis With Learned Gradient Descent
John Flynn, Michael Broxton, Paul Debevec, Matthew DuVall, Graham Fyffe, Ryan Overbeck, Noah Snavely, Richard Tucker; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2367-2376
We present a novel approach to view synthesis using multiplane images (MPIs). Building on recent advances in learned gradient descent, our algorithm generates an MPI from a set of sparse camera viewpoints. The resulting method incorporates occlusion reasoning, improving performance on challenging scene features such as object boundaries, lighting reflections, thin structures, and scenes with high depth complexity. We show that our method achieves high-quality, state-of-the-art results on two datasets: the Kalantari light field dataset, and a new camera array dataset, Spaces, which we make publicly available.
****

Animating Arbitrary Objects via Deep Motion Transfer
Aliaksandr Siarohin, Stephane Lathuiliere, Sergey Tulyakov, Elisa Ricci, Nicu Sebe; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2377-2386
This paper introduces a novel deep learning framework for image animation. Given an input image with a target object and a driving video sequence depicting a moving object, our framework generates a video in which the target object is animated according to the driving sequence. This is achieved through a deep architecture that decouples appearance and motion information. Our framework consists of three main modules: (i) a Keypoint Detector unsupervisely trained to extract object keypoints, (ii) a Dense Motion prediction network for generating dense heatmaps from sparse keypoints, in order to better encode motion information and (iii) a Motion Transfer Network, which uses the motion heatmaps and appearance information extracted from the input image to synthesize the output frames. We demonstrate the effectiveness of our method on several benchmark datasets, spanning a wide variety of object appearances, and show that our approach outperforms state-of-the-art image animation and video generation methods.
****

Textured Neural Avatars
Aliaksandra Shysheya, Egor Zakharov, Kara-Ali Aliev, Renat Bashirov, Egor Burkov, Karim Iskakov, Aleksei Ivakhnenko, Yury Malkov, Igor Pasechnik, Dmitry Ulyanov, Alexander Vakhitov, Victor Lempitsky; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2387-2397
We present a system for learning full body neural avatars, i.e. deep networks that produce full body renderings of a person for varying body pose and varying camera pose. Our system takes the middle path between the classical graphics pipeline and the recent deep learning approaches that generate images of humans using image-to-image translation. In particular, our system estimates an explicit two-dimensional texture map of the model surface. At the same time, it abstains from explicit shape modeling in 3D. Instead, at test time, the system uses a fully-convolutional network to directly map the configuration of body feature points w.r.t. the camera to the 2D texture coordinates of individual pixels in the image frame. We show that such system is capable of learning to generate realistic renderings while being trained on videos annotated with 3D poses and foreground masks. We also demonstrate that maintaining an explicit texture representation helps our system to achieve better generalization compared to systems that use direct image-to-image translation.
****

IM-Net for High Resolution Video Frame Interpolation
Tomer Peleg, Pablo Szekely, Doron Sabo, Omry Sendik; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2398-2407
Video frame interpolation is a long-studied problem in the video processing field. Recently, deep learning approaches have been applied to this problem, showing impressive results on low-resolution benchmarks. However, these methods do not scale-up favorably to high resolutions. Specifically, when the motion exceeds a typical number of pixels, their interpolation quality is degraded. Moreover, the

ir run time renders them impractical for real-time applications. In this paper w e propose IM-Net: an interpolated motion neural network. We use an economic stru ctured architecture and end-to-end training with multi-scale tailored losses. In particular, we formulate interpolated motion estimation as classification rathe r than regression. IM-Net outperforms previous methods by more than 1.3dB (PSNR) on a high resolution version of the recently introduced Vimeo triplet dataset. Moreover, the network runs in less than 33msec on a single GPU for HD resolution .

****

Homomorphic Latent Space Interpolation for Unpaired Image-To-Image Translation

Ying-Cong Chen, Xiaogang Xu, Zhuotao Tian, Jiaya Jia; Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 240 8-2416

Generative adversarial networks have achieved great success in unpaired image-to -image translation. Cycle consistency allows modeling the relationship between t wo distinct domains without paired data. In this paper, we propose an alternativ e framework, as an extension of latent space interpolation, to consider the inte rmediate region between two domains during translation. It is based on the fact that in a flat and smooth latent space, there exist many paths that connect two sample points. Properly selecting paths makes it possible to change only certain image attributes, which is useful for generating intermediate images between th e two domains. We also show that this framework can be applied to multi-domain a nd multi-modal translation. Extensive experiments manifest its generality and ap plicability to various tasks.

****

Multi-Channel Attention Selection GAN With Cascaded Semantic Guidance for Cross-View Image Translation

Hao Tang, Dan Xu, Nicu Sebe, Yanzhi Wang, Jason J. Corso, Yan Yan; Proceedi ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 2019, pp. 2417-2426

Cross-view image translation is challenging because it involves images with dras tically different views and severe deformation. In this paper, we propose a nove l approach named Multi-Channel Attention SelectionGAN (SelectionGAN) that makes it possible to generate images of natural scenes in arbitrary viewpoints, based on an image of the scene and a novel semantic map. The proposed SelectionGAN exp licitly utilizes the semantic information and consists of two stages. In the fir st stage, the condition image and the target semantic map are fed into a cycled semantic-guided generation network to produce initial coarse results. In the sec ond stage, we refine the initial results by using a multi-channel attention sele ction mechanism. Moreover, uncertainty maps automatically learned from attention s are used to guide the pixel loss for better network optimization. Extensive ex periments on Dayton, CVUSA and Ego2Top datasets show that our model is able to g enerate significantly better results than the state-of-the-art methods. The sour ce code, data and trained models are available at https://github.com/Ha0Tang/Sel ectionGAN.

****

Geometry-Consistent Generative Adversarial Networks for One-Sided Unsupervised D omain Mapping

Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, Kun Zhang, Dach eng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern R ecognition (CVPR), 2019, pp. 2427-2436

Unsupervised domain mapping aims to learn a function GXY to translate domain X t o Y in the absence of paired examples. Finding the optimal GXY without paired da ta is an ill-posed problem, so appropriate constraints are required to obtain re asonable solutions. While some prominent constraints such as cycle consistency a nd distance preservation successfully constrain the solution space, they overloo k the special properties of images that simple geometric transformations do not change the image's semantic structure. Based on this special property, we develo p a geometry-consistent generative adversarial network (Gc-GAN), which enables o ne-sided unsupervised domain mapping. GcGAN takes the original image and its cou

nterpart image transformed by a predefined geometric transformation as inputs an d generates two images in the new domain coupled with the corresponding geometry -consistency constraint. The geometry-consistency constraint reduces the space o f possible solutions while keep the correct solutions in the search space. Quant itative and qualitative comparisons with the baseline (GAN alone) and the state- of-the-art methods including CycleGAN [66] and DistanceGAN [5] demonstrate the e ffectiveness of our method.

****

DeepVoxels: Learning Persistent 3D Feature Embeddings

Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Niessner, Gordon Wetzs tein, Michael Zollhofer; Proceedings of the IEEE/CVF Conference on Computer Vis ion and Pattern Recognition (CVPR), 2019, pp. 2437-2446

In this work, we address the lack of 3D understanding of generative neural netwo rks by introducing a persistent 3D feature embedding for view synthesis. To this end, we propose DeepVoxels, a learned representation that encodes the view-depe ndent appearance of a 3D scene without having to explicitly model its geometry. At its core, our approach is based on a Cartesian 3D grid of persistent embedded features that learn to make use of the underlying 3D scene structure. Our appro ach combines insights from 3D geometric computer vision with recent advances in learning image-to-image mappings based on adversarial loss functions. DeepVoxels is supervised, without requiring a 3D reconstruction of the scene, using a 2D r e-rendering loss and enforces perspective and multi-view geometry in a principle d manner. We apply our persistent 3D scene representation to the problem of nove l view synthesis demonstrating high-quality results for a variety of challenging scenes.

****

Inverse Path Tracing for Joint Material and Lighting Estimation

Dejan Azinovic, Tzu-Mao Li, Anton Kaplanyan, Matthias Niessner; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 201 9, pp. 2447-2456

Modern computer vision algorithms have brought significant advancement to 3D geo metry reconstruction. However, illumination and material reconstruction remain l ess studied, with current approaches assuming very simplified models for materia ls and illumination. We introduce Inverse Path Tracing, a novel approach to join tly estimate the material properties of objects and light sources in indoor scen es by using an invertible light transport simulation. We assume a coarse geometr y scan, along with corresponding images and camera poses. The key contribution o f this work is an accurate and simultaneous retrieval of light sources and physi cally based material properties (e.g., diffuse reflectance, specular reflectance , roughness, etc.) for the purpose of editing and re-rendering the scene under n ew conditions. To this end, we introduce a novel optimization method using a dif ferentiable Monte Carlo renderer that computes derivatives with respect to the e stimated unknown illumination and material properties. This enables joint optimi zation for physically correct light transport and material models using a tailor ed stochastic gradient descent.

****

The Visual Centrifuge: Model-Free Layered Video Representations

Jean-Baptiste Alayrac, Joao Carreira, Andrew Zisserman; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 24 57-2466

True video understanding requires making sense of non-lambertian scenes where th e color of light arriving at the camera sensor encodes information about not jus t the last object it collided with, but about multiple mediums -- colored window s, dirty mirrors, smoke or rain. Layered video representations have the potentia l of accurately modelling realistic scenes but have so far required stringent as sumptions on motion, lighting and shape. Here we propose a learning-based approa ch for multi-layered video representation: we introduce novel uncertainty-captur ing 3D convolutional architectures and train them to separate blended videos. We show that these models then generalize to single videos, where they exhibit int eresting abilities: color constancy, factoring out shadows and separating reflec

tions. We present quantitative and qualitative results on real world videos.
****
Label-Noise Robust Generative Adversarial Networks
Takuhiro Kaneko, Yoshitaka Ushiku, Tatsuya Harada; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2467-2476
Generative adversarial networks (GANs) are a framework that learns a generative distribution through adversarial training. Recently, their class conditional extensions (e.g., conditional GAN (cGAN) and auxiliary classifier GAN (AC-GAN)) have attracted much attention owing to their ability to learn the disentangled representations and to improve the training stability. However, their training requires the availability of large-scale accurate class-labeled data, which are often laborious or impractical to collect in a real-world scenario. To remedy this, we propose a novel family of GANs called label-noise robust GANs (rGANs), which, by incorporating a noise transition model, can learn a clean label conditional generative distribution even when training labels are noisy. In particular, we propose two variants: rAC-GAN, which is a bridging model between AC-GAN and the label-noise robust classification model, and rcGAN, which is an extension of cGAN and solves this problem with no reliance on any classifier. In addition to providing the theoretical background, we demonstrate the effectiveness of our models through extensive experiments using diverse GAN configurations, various noise settings, and multiple evaluation metrics (in which we tested 402 conditions in total).
****
DLOW: Domain Flow for Adaptation and Generalization
Rui Gong, Wen Li, Yuhua Chen, Luc Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2477-2486
In this work, we present a domain flow generation(DLOW) model to bridge two different domains by generating a continuous sequence of intermediate domains flowing from one domain to the other. The benefits of our DLOW model are two-fold. First, it is able to transfer source images into different styles in the intermediate domains. The transferred images smoothly bridge the gap between source and target domains, thus easing the domain adaptation task. Second, when multiple target domains are provided for training, our DLOW model is also able to generate new styles of images that are unseen in the training data. We implement our DLOW model based on CycleGAN. A domainness variable is introduced to guide the model to generate the desired intermediate domain images. In the inference phase, a flow of various styles of images can be obtained by varying the domainness variable. We demonstrate the effectiveness of our model for both cross-domain semantic segmentation and the style generalization tasks on benchmark datasets. Our implementation is available at https://github.com/ETHRuiGong/DLOW .
****
CollaGAN: Collaborative GAN for Missing Image Data Imputation
Dongwook Lee, Junyoung Kim, Won-Jin Moon, Jong Chul Ye; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2487-2496
In many applications requiring multiple inputs to obtain a desired output, if any of the input data is missing, it often introduces large amounts of bias. Although many techniques have been developed for imputing missing data, the image imputation is still difficult due to complicated nature of natural images. To address this problem, here we proposed a novel framework for missing image data imputation, called Collaborative Generative Adversarial Network (CollaGAN). CollaGAN convert the image imputation problem to a multi-domain images-to-image translation task so that a single generator and discriminator network can successfully estimate the missing data using the remaining clean data set. We demonstrate that CollaGAN produces the images with a higher visual quality compared to the existing competing approaches in various image imputation tasks.
****
d-SNE: Domain Adaptation Using Stochastic Neighborhood Embedding
Xiang Xu, Xiong Zhou, Ragav Venkatesan, Gurumurthy Swaminathan, Orchid Majum

der; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog nition (CVPR), 2019, pp. 2497-2506
On the one hand, deep neural networks are effective in learning large datasets. On the other, they are inefficient with their data usage. They often require cop ious amount of labeled-data to train their scads of parameters. Training larger and deeper networks is hard without appropriate regularization, particularly whi le using a small dataset. Laterally, collecting well-annotated data is expensive , time-consuming and often infeasible. A popular way to regularize these network s is to simply train the network with more data from an alternate representative dataset. This can lead to adverse effects if the statistics of the representati ve dataset are dissimilar to our target.This predicament is due to the problem o f domain shift. Data from a shifted domain might not produce bespoke features wh en a feature extractor from the representative domain is used. Several techniq ues of domain adaptation have been proposed in the past to solve this problem. I n this paper, we propose a new technique (d-SNE) of domain adaptation that clev erly uses stochastic neighborhood embedding techniques and a novel modified-Haus dorff distance. The proposed technique is learnable end-to-end and is therefore, ideally suited to train neural networks. Extensive experiments demonstrate that d-SNE outperforms the current states-of-the-art and is robust to the variances in different datasets, even in the one-shot and semi-supervised learning setting s. d-SNE also demonstrates the ability to generalize to multiple domains concurr ently.
****

Taking a Closer Look at Domain Shift: Category-Level Adversaries for Semantics C onsistent Domain Adaptation
Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, Yi Yang; Proceedings of the IE EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2 507-2516
We consider the problem of unsupervised domain adaptation in semantic segmentati on. The key in this campaign consists in reducing the domain shift, i.e., enforc ing the data distributions of the two domains to be similar. A popular strategy is to align the marginal distribution in the feature space through adversarial l earning. However, this global alignment strategy does not consider the local cat egory-level feature distribution. A possible consequence of the global movement is that some categories which are originally well aligned between the source and target may be incorrectly mapped. To address this problem, this paper introduce s a category-level adversarial network, aiming to enforce local semantic consist ency during the trend of global alignment. Our idea is to take a close look at t he category-level data distribution and align each class with an adaptive advers arial loss. Specifically, we reduce the weight of the adversarial loss for categ ory-level aligned features while increasing the adversarial force for those poor ly aligned. In this process, we decide how well a feature is category-level alig ned between source and target by a co-training approach. In two domain adaptatio n tasks, i.e., GTA5 -> Cityscapes and SYNTHIA -> Cityscapes, we validate that th e proposed method matches the state of the art in segmentation accuracy.
****

ADVENT: Adversarial Entropy Minimization for Domain Adaptation in Semantic Segme ntation
Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, Patrick Perez; Pr oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2517-2526
Semantic segmentation is a key problem for many computer vision tasks. While app roaches based on convolutional neural networks constantly break new records on d ifferent benchmarks, generalizing well to diverse testing environments remains a major challenge. In numerous real-world applications, there is indeed a large g ap between data distributions in train and test domains, which results in severe performance loss at run-time. In this work, we address the task of unsupervised domain adaptation in semantic segmentation with losses based on the entropy of the pixel-wise predictions. To this end, we propose two novel, complementary met hods using (i) entropy loss and (ii) adversarial loss respectively. We demonstra

te state-of-the-art performance in semantic segmentation on two challenging "synthetic-2-real" set-ups and show that the approach can also be used for detection.

****

ContextDesc: Local Descriptor Augmentation With Cross-Modality Context

Zixin Luo, Tianwei Shen, Lei Zhou, Jiahui Zhang, Yao Yao, Shiwei Li, Tian Fang, Long Quan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2527-2536

Most existing studies on learning local features focus on the patch-based descriptions of individual keypoints, whereas neglecting the spatial relations established from their keypoint locations. In this paper, we go beyond the local detail representation by introducing context awareness to augment off-the-shelf local feature descriptors. Specifically, we propose a unified learning framework that leverages and aggregates the cross-modality contextual information, including (i) visual context from high-level image representation, and (ii) geometric context from 2D keypoint distribution. Moreover, we propose an effective N-pair loss that eschews the empirical hyper-parameter search and improves the convergence. The proposed augmentation scheme is lightweight compared with the raw local feature description, meanwhile improves remarkably on several large-scale benchmarks with diversified scenes, which demonstrates both strong practicality and generalization ability in geometric matching applications.

****

Large-Scale Long-Tailed Recognition in an Open World

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, Stella X. Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2537-2546

Real world data often have a long-tailed and open-ended distribution. A practical recognition system must classify among majority and minority classes, generalize from a few known instances, and acknowledge novelty upon a never seen instance. We define Open Long-Tailed Recognition (OLTR) as learning from such naturally distributed data and optimizing the classification accuracy over a balanced test set which include head, tail, and open classes. OLTR must handle imbalanced classification, few-shot learning, and open-set recognition in one integrated algorithm, whereas existing classification approaches focus only on one aspect and deliver poorly over the entire class spectrum. The key challenges are how to share visual knowledge between head and tail classes and how to reduce confusion between tail and open classes. We develop an integrated OLTR algorithm that maps an image to a feature space such that visual concepts can easily relate to each other based on a learned metric that respects the closed-world classification while acknowledging the novelty of the open world. Our so-called dynamic meta-embedding combines a direct image feature and an associated memory feature, with the feature norm indicating the familiarity to known classes. On three large-scale OLTR datasets we curate from object-centric ImageNet, scene-centric Places, and face-centric MS1M data, our method consistently outperforms the state-of-the-art. Our code, datasets, and models enable future OLTR research and are publicly available at https://liuziwei7.github.io/projects/LongTail.html.

****

AET vs. AED: Unsupervised Representation Learning by Auto-Encoding Transformations Rather Than Data

Liheng Zhang, Guo-Jun Qi, Liqiang Wang, Jiebo Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2547-2555

The success of deep neural networks often relies on a large amount of labeled examples, which can be difficult to obtain in many real scenarios. To address this challenge, unsupervised methods are strongly preferred for training neural networks without using any labeled data. In this paper, we present a novel paradigm of unsupervised representation learning by Auto-Encoding Transformation (AET) in contrast to the conventional Auto-Encoding Data (AED) approach. Given a randomly sampled transformation, AET seeks to predict it merely from the encoded features as accurately as possible at the output end. The idea is the following: as lo

ng as the unsupervised features successfully encode the essential information ab
out the visual structures of original and transformed images, the transformation
 can be well predicted. We will show that this AET paradigm allows us to instant
iate a large variety of transformations, from parameterized, to non-parameterize
d and GAN-induced ones. Our experiments show that AET greatly improves over exis
ting unsupervised approaches, setting new state-of-the-art performances being gr
eatly closer to the upper bounds by their fully supervised counterparts on CIFAR
-10, ImageNet and Places datasets.
****

SDC - Stacked Dilated Convolution: A Unified Descriptor Network for Dense Matchi
ng Tasks
Rene Schuster, Oliver Wasenmuller, Christian Unger, Didier Stricker; Proceedi
ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
, 2019, pp. 2556-2565
Dense pixel matching is important for many computer vision tasks such as dispari
ty and flow estimation. We present a robust, unified descriptor network that co
nsiders a large context region with high spatial variance. Our network has a ver
y large receptive field and avoids striding layers to maintain spatial resolutio
n. These properties are achieved by creating a novel neural network layer that c
onsists of multiple, parallel, stacked dilated convolutions (SDC). Several of th
ese layers are combined to form our SDC descriptor network. In our experiments,
we show that our SDC features outperform state-of-the-art feature descriptors in
 terms of accuracy and robustness. In addition, we demonstrate the superior perf
ormance of SDC in state-of-the-art stereo matching, optical flow and scene flow
algorithms on several famous public benchmarks.
****

Learning Correspondence From the Cycle-Consistency of Time
Xiaolong Wang, Allan Jabri, Alexei A. Efros; Proceedings of the IEEE/CVF Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2566-2576
We introduce a self-supervised method for learning visual correspondence from un
labeled video. The main idea is to use cycle-consistency in time as free supervi
sory signal for learning visual representations from scratch. At training time,
our model learns a feature map representation to be useful for performing cycle-
consistent tracking. At test time, we use the acquired representation to find ne
arest neighbors across space and time. We demonstrate the generalizability of th
e representation -- without finetuning -- across a range of visual correspondenc
e tasks, including video object segmentation, keypoint tracking, and optical flo
w. Our approach outperforms previous self-supervised methods and performs compet
itively with strongly supervised methods.
****

AE2-Nets: Autoencoder in Autoencoder Networks
Changqing Zhang, Yeqing Liu, Huazhu Fu; Proceedings of the IEEE/CVF Conference
 on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2577-2585
Learning on data represented with multiple views (e.g., multiple types of descri
ptors or modalities) is a rapidly growing direction in machine learning and comp
uter vision. Although effectiveness achieved, most existing algorithms usually f
ocus on classification or clustering tasks. Differently, in this paper, we focus
 on unsupervised representation learning and propose a novel framework termed Au
toencoder in Autoencoder Networks (AE^2-Nets), which integrates information from
 heterogeneous sources into an intact representation by the nested autoencoder f
ramework. The proposed method has the following merits: (1) our model jointly pe
rforms view-specific representation learning (with the inner autoencoder network
s) and multi-view information encoding (with the outer autoencoder networks) in
a unified framework; (2) due to the degradation process from the latent represen
tation to each single view, our model flexibly balances the complementarity and
consistence among multiple views. The proposed model is efficiently solved by th
e alternating direction method (ADM), and demonstrates the effectiveness compare
d with state-of-the-art algorithms.
****

Mitigating Information Leakage in Image Representations: A Maximum Entropy Appro

ach
Proteek Chandan Roy, Vishnu Naresh Boddeti; Proceedings of the IEEE/CVF Confere nce on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2586-2594
Image recognition systems have demonstrated tremendous progress over the past fe w decades thanks, in part, to our ability of learning compact and robust represe ntations of images. As we witness the wide spread adoption of these systems, it is imperative to consider the problem of unintended leakage of information from an image representation, which might compromise the privacy of the data owner. T his paper investigates the problem of learning an image representation that mini mizes such leakage of user information. We formulate the problem as an adversari al non-zero sum game of finding a good embedding function with two competing goa ls: to retain as much task dependent discriminative image information as possibl e, while simultaneously minimizing the amount of information, as measured by ent ropy, about other sensitive attributes of the user. We analyze the stability and convergence dynamics of the proposed formulation using tools from non-linear sy stems theory and compare to that of the corresponding adversarial zero-sum game formulation that optimizes likelihood as a measure of information content. Numer ical experiments on UCI, Extended Yale B, CIFAR-10 and CIFAR-100 datasets indica te that our proposed approach is able to learn image representations that exhibi t high task performance while mitigating leakage of predefined sensitive informa tion.
****

Learning Spatial Common Sense With Geometry-Aware Recurrent Networks
Hsiao-Yu Fish Tung, Ricson Cheng, Katerina Fragkiadaki; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 25 95-2603
We integrate two powerful ideas, geometry and deep visual representation learnin g, into recurrent network architectures for mobile visual scene understanding. T he proposed networks learn to "lift" 2D visual features and integrate them over time into latent 3D feature maps of the scene. They are equipped with differenti able geometric operations, such as projection, unprojection, egomotion estimatio n and stabilization, in order to compute a geometrically-consistent mapping betw een the world scene and their 3D latent feature space. We train the proposed arc hitectures to predict novel image views given short frame sequences as input. Th eir predictions strongly generalize to scenes with a novel number of objects, ap pearances and configurations, and greatly outperform predictions of previous wor ks that do not consider egomotion stabilization or a space-aware latent feature space. We train the proposed architectures to detect and segment objects in 3D, using the latent 3D feature map as input--as opposed to 2D feature maps computed from video frames. The resulting detections are permanent: they continue to exi st even when an object gets occluded or leaves the field of view. Our experiment s suggest the proposed space-aware latent feature arrangement and egomotion-stab ilized convolutions are essential architectural choices for spatial common sense to emerge in artificial embodied visual agents.
****

Structured Knowledge Distillation for Semantic Segmentation
Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, Jingdong Wang; Pr oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2604-2613
In this paper, we investigate the issue of knowledge distillation for training c ompact semantic segmentation networks by making use of cumbersome networks. We s tart from the straightforward scheme, pixel-wise distillation, which applies the distillation scheme originally introduced for image classification and performs knowledge distillation for each pixel separately. We further propose to distill the structured knowledge from cumbersome networks into compact networks, which is motivated by the fact that semantic segmentation is a structured prediction problem. We study two such structured distillation schemes: (i) pair-wise distil lation that distills the pairwise similarities, and (ii) holistic distillation t hat uses adversarial training to distill holistic knowledge. The effectiveness o f our knowledge distillation approaches is demonstrated by extensive experiment

s on three scene parsing datasets: Cityscapes, Camvid and ADE20K.
****

Scan2CAD: Learning CAD Model Alignment in RGB-D Scans
Armen Avetisyan, Manuel Dahnert, Angela Dai, Manolis Savva, Angel X. Chang, Matthias Niessner; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2614-2623
We present Scan2CAD, a novel data-driven method that learns to align clean 3D CAD models from a shape database to the noisy and incomplete geometry of a commodity RGB-D scan. For a 3D reconstruction of an indoor scene, our method takes as input a set of CAD models, and predicts a 9DoF pose that aligns each model to the underlying scan geometry. To tackle this problem, we create a new scan-to-CAD alignment dataset based on 1506 ScanNet scans with 97607 annotated keypoint pairs between 14225 CAD models from ShapeNet and their counterpart objects in the scans. Our method selects a set of representative keypoints in a 3D scan for which we find correspondences to the CAD geometry. To this end, we design a novel 3D CNN architecture that learns a joint embedding between real and synthetic objects, and from this predicts a correspondence heatmap. Based on these correspondence heatmaps, we formulate a variational energy minimization that aligns a given set of CAD models to the reconstruction. We evaluate our approach on our newly introduced Scan2CAD benchmark where we outperform both handcrafted feature descriptor as well as state-of-the-art CNN based methods by 21.39%.
****

Towards Scene Understanding: Unsupervised Monocular Depth Estimation With Semantic-Aware Representation
Po-Yi Chen, Alexander H. Liu, Yen-Cheng Liu, Yu-Chiang Frank Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2624-2632
Monocular depth estimation is a challenging task in scene understanding, with the goal to acquire the geometric properties of 3D space from 2D images. Due to the lack of RGB-depth image pairs, unsupervised learning methods aim at deriving depth information with alternative supervision such as stereo pairs. However, most existing works fail to model the geometric structure of objects, which generally results from considering pixel-level objective functions during training. In this paper, we propose SceneNet to overcome this limitation with the aid of semantic understanding from segmentation. Moreover, our proposed model is able to perform region-aware depth estimation by enforcing semantics consistency between stereo pairs. In our experiments, we qualitatively and quantitatively verify the effectiveness and robustness of our model, which produces favorable results against the state-of-the-art approaches do.
****

Tell Me Where I Am: Object-Level Scene Context Prediction
Xiaotian Qiao, Quanlong Zheng, Ying Cao, Rynson W.H. Lau; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2633-2641
Contextual information has been shown to be effective in helping solve various image understanding tasks. Previous works have focused on the extraction of contextual information from an image and use it to infer the properties of some object(s) in the image. In this paper, we consider an inverse problem of how to hallucinate missing contextual information from the properties of a few standalone objects. We refer to it as scene context prediction. This problem is difficult as it requires an extensive knowledge of complex and diverse relationships among different objects in natural scenes. We propose a convolutional neural network, which takes as input the properties (i.e., category, shape, and position) of a few standalone objects to predict an object-level scene layout that compactly encodes the semantics and structure of the scene context where the given objects are. Our quantitative experiments and user studies show that our model can generate more plausible scene context than the baseline approach. We demonstrate that our model allows for the synthesis of realistic scene images from just partial scene layouts and internally learns useful features for scene recognition.
****

Normalized Object Coordinate Space for Category-Level 6D Object Pose and Size Estimation
He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, Leonidas J. Guibas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2642-2651
The goal of this paper is to estimate the 6D pose and dimensions of unseen object instances in an RGB-D image. Contrary to "instance-level" 6D pose estimation tasks, our problem assumes that no exact object CAD models are available during either training or testing time. To handle different and unseen object instances in a given category, we introduce a Normalized Object Coordinate Space (NOCS)---a shared canonical representation for all possible object instances within a category. Our region-based neural network is then trained to directly infer the correspondence from observed pixels to this shared object representation (NOCS) along with other object information such as class label and instance mask. These predictions can be combined with the depth map to jointly estimate the metric 6D pose and dimensions of multiple objects in a cluttered scene. To train our network, we present a new context-aware technique to generate large amounts of fully annotated mixed reality data. To further improve our model and evaluate its performance on real data, we also provide a fully annotated real-world dataset with large environment and instance variation. Extensive experiments demonstrate that the proposed method is able to robustly estimate the pose and size of unseen object instances in real environments while also achieving state-of-the-art performance on standard 6D pose estimation benchmarks.
****
Supervised Fitting of Geometric Primitives to 3D Point Clouds
Lingxiao Li, Minhyuk Sung, Anastasia Dubrovina, Li Yi, Leonidas J. Guibas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2652-2660
Fitting geometric primitives to 3D point cloud data bridges a gap between low-level digitized 3D data and high-level structural information on the underlying 3D shapes. As such, it enables many downstream applications in 3D data processing. For a long time, RANSAC-based methods have been the gold standard for such primitive fitting problems, but they require careful per-input parameter tuning and thus do not scale well for large datasets with diverse shapes. In this work, we introduce Supervised Primitive Fitting Network (SPFN), an end-to-end neural network that can robustly detect a varying number of primitives at different scales without any user control. The network is supervised using ground truth primitive surfaces and primitive membership for the input points. Instead of directly predicting the primitives, our architecture first predicts per-point properties and then uses a differential model estimation module to compute the primitive type and parameters. We evaluate our approach on a novel benchmark of ANSI 3D mechanical component models and demonstrate a significant improvement over both the state-of-the-art RANSAC-based methods and the direct neural prediction.
****
Do Better ImageNet Models Transfer Better?
Simon Kornblith, Jonathon Shlens, Quoc V. Le; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2661-2671
Transfer learning is a cornerstone of computer vision, yet little work has been done to evaluate the relationship between architecture and transfer. An implicit hypothesis in modern computer vision research is that models that perform better on ImageNet necessarily perform better on other vision tasks. However, this hypothesis has never been systematically tested. Here, we compare the performance of 16 classification networks on 12 image classification datasets. We find that, when networks are used as fixed feature extractors or fine-tuned, there is a strong correlation between ImageNet accuracy and transfer accuracy (r = 0.99 and 0.96, respectively). In the former setting, we find that this relationship is very sensitive to the way in which networks are trained on ImageNet; many common forms of regularization slightly improve ImageNet accuracy but yield features that are much worse for transfer learning. Additionally, we find that, on two small fine-grained image classification datasets, pretraining on ImageNet provides min

imal benefits, indicating the learned features from ImageNet do not transfer wel
l to fine-grained tasks. Together, our results show that ImageNet architectures
generalize well across datasets, but ImageNet features are less general than pre
viously suggested.
****

Gotta Adapt 'Em All: Joint Pixel and Feature-Level Domain Adaptation for Recogni
tion in the Wild

Luan Tran, Kihyuk Sohn, Xiang Yu, Xiaoming Liu, Manmohan Chandraker; Proceed
ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR
), 2019, pp. 2672-2681

Recent developments in deep domain adaptation have allowed knowledge transfer fr
om a labeled source domain to an unlabeled target domain at the level of interme
diate features or input pixels. We propose that advantages may be derived by com
bining them, in the form of different insights that lead to a novel design and c
omplementary properties that result in better performance. At the feature level,
inspired by insights from semi-supervised learning, we propose a classification
-aware domain adversarial neural network that brings target examples into more c
lassifiable regions of source domain. Next, we posit that computer vision insigh
ts are more amenable to injection at the pixel level. In particular, we use 3D g
eometry and image synthesis based on a generalized appearance flow to preserve i
dentity across pose transformations, while using an attribute-conditioned CycleG
AN to translate a single source into multiple target images that differ in lower
-level properties such as lighting. Besides standard UDA benchmark, we validate
on a novel and apt problem of car recognition in unlabeled surveillance images u
sing labeled images from the web, handling explicitly specified, nameable factor
s of variation through pixel-level and implicit, unspecified factors through fea
ture-level adaptation.
****

Understanding the Disharmony Between Dropout and Batch Normalization by Variance
Shift

Xiang Li, Shuo Chen, Xiaolin Hu, Jian Yang; Proceedings of the IEEE/CVF Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2682-2690

This paper first answers the question "why do the two most powerful techniques D
ropout and Batch Normalization (BN) often lead to a worse performance when they
are combined together in many modern neural networks, but cooperate well sometim
es as in Wide ResNet (WRN)?" in both theoretical and empirical aspects. Theoreti
cally, we find that Dropout shifts the variance of a specific neural unit when w
e transfer the state of that network from training to test. However, BN maintain
s its statistical variance, which is accumulated from the entire learning proced
ure, in the test phase. The inconsistency of variances in Dropout and BN (we na
me this scheme "variance shift") causes the unstable numerical behavior in infer
ence that leads to erroneous predictions finally. Meanwhile, the large feature d
imension in WRN further reduces the "variance shift" to bring benefits to the ov
erall performance. Thorough experiments on representative modern convolutional n
etworks like DenseNet, ResNet, ResNeXt and Wide ResNet confirm our findings. Acc
ording to the uncovered mechanism, we get better understandings in the combinati
on of these two techniques and summarize guidelines for better practices.
****

Circulant Binary Convolutional Networks: Enhancing the Performance of 1-Bit DCNN
s With Circulant Back Propagation

Chunlei Liu, Wenrui Ding, Xin Xia, Baochang Zhang, Jiaxin Gu, Jianzhuang Li
u, Rongrong Ji, David Doermann; Proceedings of the IEEE/CVF Conference on Comp
uter Vision and Pattern Recognition (CVPR), 2019, pp. 2691-2699

The rapidly decreasing computation and memory cost has recently driven the succe
ss of many applications in the field of deep learning. Practical applications of
deep learning in resource-limited hardware, such as embedded devices and smart
phones, however, remain challenging. For binary convolutional networks, the reas
on lies in the degraded representation caused by binarizing full-precision filte
rs. To address this problem, we propose new circulant filters (CiFs) and a circu
lant binary convolution (CBConv) to enhance the capacity of binarized convolutio

nal features via our circulant back propagation (CBP). The CiFs can be easily in corporated into existing deep convolutional neural networks (DCNNs), which leads to new Circulant Binary Convolutional Networks (CBCNs). Extensive experiments c onfirm that the performance gap between the 1-bit and full-precision DCNNs is mi nimized by increasing the filter diversity, which further increases the represen tational ability in our networks. Our experiments on ImageNet show that CBCNs ac hieve 61.4% top-1 accuracy with ResNet18. Compared to the state-of-the-art such as XNOR, CBCNs can achieve up to 10% higher top-1 accuracy with more powerful re presentational ability.

****

DeFusionNET: Defocus Blur Detection via Recurrently Fusing and Refining Multi-Sc ale Deep Features

Chang Tang, Xinzhong Zhu, Xinwang Liu, Lizhe Wang, Albert Zomaya; Proceeding s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2700-2709

Defocus blur detection aims to detect out-of-focus regions from an image. Althou gh attracting more and more attention due to its widespread applications, defocu s blur detection still confronts several challenges such as the interference of background clutter, sensitivity to scales and missing boundary details of defocu s blur regions. To deal with these issues, we propose a deep neural network whic h recurrently fuses and refines multi-scale deep features (DeFusionNet) for defo cus blur detection. We firstly utilize a fully convolutional network to extract multi-scale deep features. The features from bottom layers are able to capture r ich low-level features for details preservation, while the features from top lay ers can characterize the semantic information to locate blur regions. These feat ures from different layers are fused as shallow features and semantic features, respectively. After that, the fused shallow features are propagated to top layer s for refining the fine details of detected defocus blur regions, and the fused semantic features are propagated to bottom layers to assist in better locating t he defocus regions. The feature fusing and refining are carried out in a recurre nt manner. Also, we finally fuse the output of each layer at the last recurrent step to obtain the final defocus blur map by considering the sensitivity to scal es of the defocus degree. Experiments on two commonly used defocus blur detectio n benchmark datasets are conducted to demonstrate the superority of DeFusionNet when compared with other 10 competitors. Code and more results can be found at: http://tangchang.net

****

Deep Virtual Networks for Memory Efficient Inference of Multiple Tasks

Eunwoo Kim, Chanho Ahn, Philip H.S. Torr, Songhwai Oh; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 27 10-2719

Deep networks consume a large amount of memory by their nature. A natural questi on arises can we reduce that memory requirement whilst maintaining performance. In particular, in this work we address the problem of memory efficient learning for multiple tasks. To this end, we propose a novel network architecture produci ng multiple networks of different configurations, termed deep virtual networks (DVNs), for different tasks. Each DVN is specialized for a single task and struct ured hierarchically. The hierarchical structure, which contains multiple levels of hierarchy corresponding to different numbers of parameters, enables multiple inference for different memory budgets. The building block of a deep virtual net work is based on a disjoint collection of parameters of a network, which we call a unit. The lowest level of hierarchy in a deep virtual network is a unit, and higher levels of hierarchy contain lower levels' units and other additional unit s. Given a budget on the number of parameters, a different level of a deep virtu al network can be chosen to perform the task. A unit can be shared by different DVNs, allowing multiple DVNs in a single network. In addition, shared units prov ide assistance to the target task with additional knowledge learned from another tasks. This cooperative configuration of DVNs makes it possible to handle diffe rent tasks in a memory-aware manner. Our experiments show that the proposed meth od outperforms existing approaches for multiple tasks. Notably, ours is more eff

icient than others as it allows memory-aware inference for all tasks.
****

Universal Domain Adaptation
Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, Michael I. Jordan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2720-2729
Domain adaptation aims to transfer knowledge in the presence of the domain gap. Existing domain adaptation methods rely on rich prior knowledge about the relationship between the label sets of source and target domains, which greatly limits their application in the wild. This paper introduces Universal Domain Adaptation (UDA) that requires no prior knowledge on the label sets. For a given source label set and a target label set, they may contain a common label set and hold a private label set respectively, bringing up an additional category gap. UDA requires a model to either (1) classify the target sample correctly if it is associated with a label in the common label set, or (2) mark it as "unknown" otherwise. More importantly, a UDA model should work stably against a wide spectrum of commonness (the proportion of the common label set over the complete label set) so that it can handle real-world problems with unknown target label sets. To solve the universal domain adaptation problem, we propose Universal Adaptation Network (UAN). It quantifies sample-level transferability to discover the common label set and the label sets private to each domain, thereby promoting the adaptation in the automatically discovered common label set and recognizing the "unknown" samples successfully. A thorough evaluation shows that UAN outperforms the state of the art closed set, partial and open set domain adaptation methods in the novel UDA setting.
****

Improving Transferability of Adversarial Examples With Input Diversity
Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, Alan L. Yuille; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2730-2739
Though CNNs have achieved the state-of-the-art performance on various vision tasks, they are vulnerable to adversarial examples --- crafted by adding human-imperceptible perturbations to clean images. However, most of the existing adversarial attacks only achieve relatively low success rates under the challenging black-box setting, where the attackers have no knowledge of the model structure and parameters. To this end, we propose to improve the transferability of adversarial examples by creating diverse input patterns. Instead of only using the original images to generate adversarial examples, our method applies random transformations to the input images at each iteration. Extensive experiments on ImageNet show that the proposed attack method can generate adversarial examples that transfer much better to different networks than existing baselines. By evaluating our method against top defense solutions and official baselines from NIPS 2017 adversarial competition, the enhanced attack reaches an average success rate of 73.0%, which outperforms the top-1 attack submission in the NIPS competition by a large margin of 6.6%. We hope that our proposed attack strategy can serve as a strong benchmark baseline for evaluating the robustness of networks to adversaries and the effectiveness of different defense methods in the future. Code is available at https://github.com/cihangxie/DI-2-FGSM.
****

Sequence-To-Sequence Domain Adaptation Network for Robust Text Image Recognition
Yaping Zhang, Shuai Nie, Wenju Liu, Xing Xu, Dongxiang Zhang, Heng Tao Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2740-2749
Domain adaptation has shown promising advances for alleviating domain shift problem. However, recent visual domain adaptation works usually focus on non-sequential object recognition with a global coarse alignment, which is inadequate to transfer effective knowledge for sequence-like text images with variable-length fine-grained character information. In this paper, we develop a Sequence-to-Sequence Domain Adaptation Network (SSDAN) for robust text image recognition, which could exploit unsupervised sequence data by an attention-based sequence encoder-de

coder network. In the SSDAN, a gated attention similarity (GAS) unit is introduced to adaptively focus on aligning the distribution of the source and target sequence data in an attended character-level feature space rather than a global coarse alignment. Extensive text recognition experiments show the SSDAN could efficiently transfer sequence knowledge and validate the promising power of the proposed model towards real world applications in various recognition scenarios, including the natural scene text, handwritten text and even mathematical expression recognition.

****

Hybrid-Attention Based Decoupled Metric Learning for Zero-Shot Image Retrieval
Binghui Chen, Weihong Deng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2750-2759

In zero-shot image retrieval (ZSIR) task, embedding learning becomes more attractive, however, many methods follow the traditional metric learning idea and omit the problems behind zero-shot settings. In this paper, we first emphasize the importance of learning visual discriminative metric and preventing the partial/selective learning behavior of learner in ZSIR, and then propose the Decoupled Metric Learning (DeML) framework to achieve these individually. Instead of coarsely optimizing an unified metric, we decouple it into multiple attention-specific parts so as to recurrently induce the discrimination and explicitly enhance the generalization. And they are mainly achieved by our object-attention module based on random walk graph propagation and the channel-attention module based on the adversary constraint, respectively. We demonstrate the necessity of addressing the vital problems in ZSIR on the popular benchmarks, outperforming the state-of-the-art methods by a significant margin. Code is available at http://www.bhchen.cn

****

Learning to Sample
Oren Dovrat, Itai Lang, Shai Avidan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2760-2769

Processing large point clouds is a challenging task. Therefore, the data is often sampled to a size that can be processed more easily. The question is how to sample the data? A popular sampling technique is Farthest Point Sampling (FPS). However, FPS is agnostic to a downstream application (classification, retrieval, etc.). The underlying assumption seems to be that minimizing the farthest point distance, as done by FPS, is a good proxy to other objective functions. We show that it is better to learn how to sample. To do that, we propose a deep network to simplify 3D point clouds. The network, termed S-NET, takes a point cloud and produces a smaller point cloud that is optimized for a particular task. The simplified point cloud is not guaranteed to be a subset of the original point cloud. Therefore, we match it to a subset of the original points in a post-processing step. We contrast our approach with FPS by experimenting on two standard data sets and show significantly better results for a variety of applications. Our code is publicly available.

****

Few-Shot Learning via Saliency-Guided Hallucination of Samples
Hongguang Zhang, Jing Zhang, Piotr Koniusz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2770-2779

Learning new concepts from a few of samples is a standard challenge in computer vision. The main directions to improve the learning ability of few-shot training models include (i) a robust similarity learning and (ii) generating or hallucinating additional data from the limited existing samples. In this paper, we follow the latter direction and present a novel data hallucination model. Currently, most datapoint generators contain a specialized network (i.e., GAN) tasked with hallucinating new datapoints, thus requiring large numbers of annotated data for their training in the first place. In this paper, we propose a novel less-costly hallucination method for few-shot learning which utilizes saliency maps. To this end, we employ a saliency network to obtain the foregrounds and backgrounds of available image samples and feed the resulting maps into a two-stream network to hallucinate datapoints directly in the feature space from viable foreground-

background combinations. To the best of our knowledge, we are the first to lever age saliency maps for such a task and we demonstrate their usefulness in halluci nating additional datapoints for few-shot learning. Our proposed network achieve s the state of the art on publicly available datasets.
****

Variational Convolutional Neural Network Pruning
Chenglong Zhao, Bingbing Ni, Jian Zhang, Qiwei Zhao, Wenjun Zhang, Qi Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti on (CVPR), 2019, pp. 2780-2789
We propose a variational Bayesian scheme for pruning convolutional neural networ ks in channel level. This idea is motivated by the fact that deterministic value based pruning methods are inherently improper and unstable. In a nutshell, var iational technique is introduced to estimate distribution of a newly proposed p arameter, called channel saliency, based on this, redundant channels can be rem oved from model via a simple criterion. The advantages are two-fold: 1) Our meth od conducts channel pruning without desire of re-training stage, thus improving the computation efficiency. 2) Our method is implemented as a stand-alone modul e, called variational pruning layer, which can be straightforwardly inserted int o off-the-shelf deep learning packages, without any special network design. Ext ensive experimental results well demonstrate the effectiveness of our method: Fo r CIFAR-10, we perform channel removal on different CNN models up to 74% reduct ion, which results in significant size reduction and computation saving. For Ima geNet, about 40% channels of ResNet-50 are removed without compromising accurac y.
****

Towards Optimal Structured CNN Pruning via Generative Adversarial Learning
Shaohui Lin, Rongrong Ji, Chenqian Yan, Baochang Zhang, Liujuan Cao, Qixian g Ye, Feiyue Huang, David Doermann; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2790-2799
Structured pruning of filters or neurons has received increased focus for compre ssing convolutional neural networks. Most existing methods rely on multi-stage o ptimizations in a layer-wise manner for iteratively pruning and retraining which may not be optimal and may be computation intensive. Besides, these methods are designed for pruning a specific structure, such as filter or block structures w ithout jointly pruning heterogeneous structures. In this paper, we propose an ef fective structured pruning approach that jointly prunes filters as well as other structures in an end-to-end manner. To accomplish this, we first introduce a so ft mask to scale the output of these structures by defining a new objective func tion with sparsity regularization to align the output of baseline and network wi th this mask. We then effectively solve the optimization problem by generative a dversarial learning (GAL), which learns a sparse soft mask in a label-free and a n end-to-end manner. By forcing more scale factors in the soft mask to zero, the fast iterative shrinkage-thresholding algorithm (FISTA) can be leveraged to fas t and reliably remove the corresponding structures. Extensive experiments demons trate the effectiveness of GAL on different datasets, including MNIST, CIFAR-10 and ImageNet ILSVRC 2012. For example, on ImageNet ILSVRC 2012, the pruned ResNe t-50 achieves 10.88% Top-5 error and results in a factor of 3.7x speedup. This s ignificantly outperforms state-of-the-art methods.
****

Exploiting Kernel Sparsity and Entropy for Interpretable CNN Compression
Yuchao Li, Shaohui Lin, Baochang Zhang, Jianzhuang Liu, David Doermann, Yon gjian Wu, Feiyue Huang, Rongrong Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2800-2809
Compressing convolutional neural networks (CNNs) has received ever-increasing re search focus. However, most existing CNN compression methods do not interpret th eir inherent structures to distinguish the implicit redundancy. In this paper, w e investigate the problem of CNN compression from a novel interpretable perspect ive. The relationship between the input feature maps and 2D kernels is revealed in a theoretical framework, based on which a kernel sparsity and entropy (KSE) i ndicator is proposed to quantitate the feature map importance in a feature-agnos

tic manner to guide model compression. Kernel clustering is further conducted ba sed on the KSE indicator to accomplish high-precision CNN compression. KSE is ca pable of simultaneously compressing each layer in an efficient way, which is sig nificantly faster compared to previous data-driven feature map pruning methods. We comprehensively evaluate the compression and speedup of the proposed method o n CIFAR-10, SVHN and ImageNet 2012. Our method demonstrates superior performance gains over previous ones. In particular, it achieves 4.7x FLOPs reduction and 2 .9x compression on ResNet-50 with only a top-5 accuracy drop of 0.35% on ImageNe t 2012, which significantly outperforms state-of-the-art methods.
****

Fully Quantized Network for Object Detection
Rundong Li, Yan Wang, Feng Liang, Hongwei Qin, Junjie Yan, Rui Fan; Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2810-2819
Efficient neural network inference is important in a number of practical domains , such as deployment in mobile settings. An effective method for increasing inf erence efficiency is to use low bitwidth arithmetic, which can subsequently be a ccelerated using dedicated hardware. However, designing effective quantization schemes while maintaining network accuracy is challenging. In particular, curre nt techniques face difficulty in performing fully end-to-end quantization, makin g use of aggressively low bitwidth regimes such as 4-bit, and applying quantized networks to complex tasks such as object detection. In this paper, we demonstr ate that many of these difficulties arise because of instability during the fine -tuning stage of the quantization process, and propose several novel techniques to overcome these instabilities. We apply our techniques to produce fully quanti zed 4-bit detectors based on RetinaNet and Faster R-CNN, and show that these ach ieve state-of-the-art performance for quantized detectors. The mAP loss due to q uantization using our methods is more than 3.8x less than the loss from existing methods.
****

MnasNet: Platform-Aware Neural Architecture Search for Mobile
Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, Quoc V. Le; Proceedings of the IEEE/CVF Conference on Computer Vision a nd Pattern Recognition (CVPR), 2019, pp. 2820-2828
Designing convolutional neural networks (CNN) for mobile devices is challenging because mobile models need to be small and fast, yet still accurate. Although si gnificant efforts have been dedicated to design and improve mobile CNNs on all d imensions, it is very difficult to manually balance these trade-offs when there are so many architectural possibilities to consider. In this paper, we propose a n automated mobile neural architecture search (MNAS) approach, which explicitly incorporate model latency into the main objective so that the search can identif y a model that achieves a good trade-off between accuracy and latency. Unlike pr evious work, where latency is considered via another, often inaccurate proxy (e. g., FLOPS), our approach directly measures real-world inference latency by execu ting the model on mobile phones. To further strike the right balance between fle xibility and search space size, we propose a novel factorized hierarchical searc h space that encourages layer diversity throughout the network. Experimental res ults show that our approach consistently outperforms state-of-the-art mobile CNN models across multiple vision tasks. On the ImageNet classification task, our M nasNet achieves 75.2% top-1 accuracy with 78ms latency on a Pixel phone, which i s 1.8x faster than MobileNetV2 with 0.5% higher accuracy and 2.3x faster than NA SNet with 1.2% higher accuracy. Our MnasNet also achieves better mAP quality tha n MobileNets for COCO object detection. Code is at https://github.com/tensorflow /tpu/tree/master/models/official/mnasnet.
****

Student Becoming the Master: Knowledge Amalgamation for Joint Scene Parsing, Dep th Estimation, and More
Jingwen Ye, Yixin Ji, Xinchao Wang, Kairi Ou, Dapeng Tao, Mingli Song; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2019, pp. 2829-2838

In this paper, we investigate a novel deep-model reusing task. Our goal is to train a lightweight and versatile student model, without human-labelled annotations, that amalgamates the knowledge and masters the expertise of two pre-trained teacher models working on heterogeneous problems, one on scene parsing and the other on depth estimation. To this end, we propose an innovative training strategy that learns the parameters of the student intertwined with the teachers, achieved by "projecting" its amalgamated features onto each teacher's domain and computing the loss. We also introduce two options to generalize the proposed training strategy to handle three or more tasks simultaneously. The proposed scheme yields very encouraging results. As demonstrated on several benchmarks, the trained student model achieves results even superior to those of the teachers in their own expertise domains and on par with the state-of-the-art fully supervised models relying on human-labelled annotations.
****

K-Nearest Neighbors Hashing

Xiangyu He, Peisong Wang, Jian Cheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2839-2848

Hashing based approximate nearest neighbor search embeds high dimensional data to compact binary codes, which enables efficient similarity search and storage. However, the non-isometry sign() function makes it hard to project the nearest neighbors in continuous data space into the closest codewords in discrete Hamming space. In this work, we revisit the sign() function from the perspective of space partitioning. In specific, we bridge the gap between k-nearest neighbors and binary hashing codes with Shannon entropy. We further propose a novel K-Nearest Neighbors Hashing (KNNH) method to learn binary representations from KNN within the subspaces generated by sign(). Theoretical and experimental results show that the KNN relation is of central importance to neighbor preserving embeddings, and the proposed method outperforms the state-of-the-arts on benchmark datasets.
****

Learning RoI Transformer for Oriented Object Detection in Aerial Images

Jian Ding, Nan Xue, Yang Long, Gui-Song Xia, Qikai Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2849-2858

Object detection in aerial images is an active yet challenging task in computer vision because of the bird's-eye view perspective, the highly complex backgrounds, and the variant appearances of objects. Especially when detecting densely packed objects in aerial images, methods relying on horizontal proposals for common object detection often introduce mismatches between the Region of Interests (RoIs) and objects. This leads to the common misalignment between the final object classification confidence and localization accuracy. In this paper, we propose a RoI Transformer to address these problems. The core idea of RoI Transformer is to apply spatial transformations on RoIs and learn the transformation parameters under the supervision of oriented bounding box (OBB) annotations. RoI Transformer is with lightweight and can be easily embedded into detectors for oriented object detection. Simply apply the RoI Transformer to light head RCNN has achieved state-of-the-art performances on two common and challenging aerial datasets, i.e., DOTA and HRSC2016, with a neglectable reduction to detection speed. Our RoI Transformer exceeds the deformable Position Sensitive RoI pooling when oriented bounding-box annotations are available. Extensive experiments have also validated the flexibility and effectiveness of our RoI Transformer.
****

Snapshot Distillation: Teacher-Student Optimization in One Generation

Chenglin Yang, Lingxi Xie, Chi Su, Alan L. Yuille; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2859-2868

Optimizing a deep neural network is a fundamental task in computer vision, yet direct training methods often suffer from over-fitting. Teacher-student optimization aims at providing complementary cues from a model trained previously, but these approaches are often considerably slow due to the pipeline of training a few generations in sequence, i.e., time complexity is increased by several times.

This paper presents snapshot distillation (SD), the first framework which enables teacher-student optimization in one generation. The idea of SD is very simple: instead of borrowing supervision signals from previous generations, we extract such information from earlier epochs in the same generation, meanwhile make sure that the difference between teacher and student is sufficiently large so as to prevent under-fitting. To achieve this goal, we implement SD in a cyclic learning rate policy, in which the last snapshot of each cycle is used as the teacher for all iterations in the next cycle, and the teacher signal is smoothed to provide richer information. In standard image classification benchmarks such as CIFAR100 and ILSVRC2012, SD achieves consistent accuracy gain without heavy computational overheads. We also verify that models pre-trained with SD transfers well to object detection and semantic segmentation in the PascalVOC dataset.
****

Geometry-Aware Distillation for Indoor Semantic Segmentation
Jianbo Jiao, Yunchao Wei, Zequn Jie, Honghui Shi, Rynson W.H. Lau, Thomas S. Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2869-2878
It has been shown that jointly reasoning the 2D appearance and 3D information from RGB-D domains is beneficial to indoor scene semantic segmentation. However, most existing approaches require accurate depth map as input to segment the scene which severely limits their applications. In this paper, we propose to jointly infer the semantic and depth information by distilling geometry-aware embedding to eliminate such strong constraint while still exploiting the helpful depth domain information. In addition, we use this learned embedding to improve the quality of semantic segmentation, through a proposed geometry-aware propagation framework followed by several multi-level skip feature fusion blocks. By decoupling the single task prediction network into two joint tasks of semantic segmentation and geometry embedding learning, together with the proposed information propagation and feature fusion architecture, our method is shown to perform favorably against state-of-the-art methods for semantic segmentation on publicly available challenging indoor datasets.
****

LiveSketch: Query Perturbations for Guided Sketch-Based Visual Search
John Collomosse, Tu Bui, Hailin Jin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2879-2887
LiveSketch is a novel algorithm for searching large image collections using hand-sketched queries. LiveSketch tackles the inherent ambiguity of sketch search by creating visual suggestions that augment the query as it is drawn, making query specification an iterative rather than one-shot process that helps disambiguate users' search intent. Our technical contributions are: a triplet convnet architecture that incorporates an RNN based variational autoencoder to search for images using vector (stroke-based) queries; real-time clustering to identify likely search intents (and so, targets within the search embedding); and the use of backpropagation from those targets to perturb the input stroke sequence, so suggesting alterations to the query in order to guide the search. We show improvements in accuracy and time-to-task over contemporary baselines using a 67M image corpus.
****

Bounding Box Regression With Uncertainty for Accurate Object Detection
Yihui He, Chenchen Zhu, Jianren Wang, Marios Savvides, Xiangyu Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2888-2897
Large-scale object detection datasets (e.g., MS-COCO) try to define the ground truth bounding boxes as clear as possible. However, we observe that ambiguities are still introduced when labeling the bounding boxes. In this paper, we propose a novel bounding box regression loss for learning bounding box transformation and localization variance together. Our loss greatly improves the localization accuracies of various architectures with nearly no additional computation. The learned localization variance allows us to merge neighboring bounding boxes during non-maximum suppression (NMS), which further improves the localization performanc

e. On MS-COCO, we boost the Average Precision (AP) of VGG-16 Faster R-CNN from 2
3.6% to 29.1%. More importantly, for ResNet-50-FPN Mask R-CNN, our method improv
es the AP and AP90 by 1.8% and 6.2% respectively, which significantly outperform
s previous state-of-the-art bounding box refinement methods. Our code and models
 are available at github.com/yihui-he/KL-Loss
****

OCGAN: One-Class Novelty Detection Using GANs With Constrained Latent Representa
tions
Pramuditha Perera, Ramesh Nallapati, Bing Xiang; Proceedings of the IEEE/CVF C
onference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2898-2906
We present a novel model called OCGAN for the classical problem of one-class nov
elty detection, where, given a set of examples from a particular class, the goal
 is to determine if a query example is from the same class. Our solution is base
d on learning latent representations of in-class examples using a de-noising aut
o-encoder network. The key contribution of our work is our proposal to explicitl
y constrain the latent space to exclusively represent the given class. In order
to accomplish this goal, firstly, we force the latent space to have bounded supp
ort by introducing a tanh activation in the encoder's output layer. Secondly, us
ing a discriminator in the latent space that is trained adversarially, we ensure
 that encoded representations of in-class examples resemble uniform random sampl
es drawn from the same bounded space. Thirdly, using a second adversarial discri
minator in the input space, we ensure all randomly drawn latent samples generate
 examples that look real. Finally, we introduce a gradient-descent based samplin
g technique that explores points in the latent space that generate potential out
-of-class examples, which are fed back to the network to further train it to gen
erate in-class examples from those points. The effectiveness of the proposed met
hod is measured across four publicly available datasets using two one-class nove
lty detection protocols where we achieve state-of-the-art results.
****

Learning Metrics From Teachers: Compact Networks for Image Embedding
Lu Yu, Vacit Oguz Yazici, Xialei Liu, Joost van de Weijer, Yongmei Cheng, A
rnau Ramisa; Proceedings of the IEEE/CVF Conference on Computer Vision and Patte
rn Recognition (CVPR), 2019, pp. 2907-2916
Metric learning networks are used to compute image embeddings, which are widely
used in many applications such as image retrieval and face recognition. In this
paper, we propose to use network distillation to efficiently compute image embed
dings with small networks. Network distillation has been successfully applied to
 improve image classification, but has hardly been explored for metric learning.
 To do so, we propose two new loss functions that model the communication of a d
eep teacher network to a small student network. We evaluate our system in severa
l datasets, including CUB-200-2011, Cars-196, Stanford Online Products and show
that embeddings computed using small student networks perform significantly bett
er than those computed using standard networks of similar size. Results on a ver
y compact network (MobileNet-0.25), which can be used on mobile devices, show th
at the proposed method can greatly improve Recall@1 results from 27.5% to 44.6%.
 Furthermore, we investigate various aspects of distillation for embeddings, inc
luding hint and attention layers, semi-supervised learning and cross quality dis
tillation. (Code is available at https://github.com/yulu0724/EmbeddingDistillati
on).
****

Activity Driven Weakly Supervised Object Detection
Zhenheng Yang, Dhruv Mahajan, Deepti Ghadiyaram, Ram Nevatia, Vignesh Ramana
than; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Reco
gnition (CVPR), 2019, pp. 2917-2926
Weakly supervised object detection aims at reducing the amount of supervision re
quired to train detection models. Such models are traditionally learned from ima
ges/videos labelled only with the object class and not the object bounding box.
In our work, we try to leverage not only the object class labels but also the ac
tion labels associated with the data. We show that the action depicted in the im
age/video can provide strong cues about the location of the associated object. W

e learn a spatial prior for the object dependent on the action (e.g. "ball" is c
loser to "leg of the person" in "kicking ball"), and incorporate this prior to s
imultaneously train a joint object detection and action classification model. We
 conducted experiments on both video datasets and image datasets to evaluate the
 performance of our weakly supervised object detection model. Our approach outpe
rformed the current state-of-the-art (SOTA) method by more than 6% in mAP on the
 Charades video dataset.
****

Separate to Adapt: Open Set Domain Adaptation via Progressive Separation
Hong Liu, Zhangjie Cao, Mingsheng Long, Jianmin Wang, Qiang Yang; Proceeding
s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
2019, pp. 2927-2936
Domain adaptation has become a resounding success in leveraging labeled data fro
m a source domain to learn an accurate classifier for an unlabeled target domain
. When deployed in the wild, the target domain usually contains unknown classes
that are not observed in the source domain. Such setting is termed Open Set Doma
in Adaptation (OSDA). While several methods have been proposed to address OSDA,
none of them takes into account the openness of the target domain, which is meas
ured by the proportion of unknown classes in all target classes. Openness is a c
ritical point in open set domain adaptation and exerts a significant impact on p
erformance. In addition, current work aligns the entire target domain with the s
ource domain without excluding unknown samples, which may give rise to negative
transfer due to the mismatch between unknown and known classes. To this end, thi
s paper presents Separate to Adapt (STA), an end-to-end approach to open set dom
ain adaptation. The approach adopts a coarse-to-fine weighting mechanism to prog
ressively separate the samples of unknown and known classes, and simultaneously
weigh their importance on feature distribution alignment. Our approach allows op
enness-robust open set domain adaptation, which can be adaptive to a variety of
openness in the target domain. We evaluate STA on several benchmark datasets of
various openness levels. Results verify that STA significantly outperforms previ
ous methods.
****

Layout-Graph Reasoning for Fashion Landmark Detection
Weijiang Yu, Xiaodan Liang, Ke Gong, Chenhan Jiang, Nong Xiao, Liang Lin; P
roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
 (CVPR), 2019, pp. 2937-2945
Detecting dense landmarks for diverse clothes, as a fundamental technique for cl
othes analysis, has attracted increasing research attention due to its huge appl
ication potential. However, due to the lack of modeling underlying semantic layo
ut constraints among landmarks, prior works often detect ambiguous and structure
-inconsistent landmarks of multiple overlapped clothes in one person. In this pa
per, we propose to seamlessly enforce structural layout relationships among land
marks on the intermediate representations via multiple stacked layout-graph reas
oning layers. We define the layout-graph as a hierarchical structure including a
 root node, body-part nodes (e.g. upper body, lower body), coarse clothes-part n
odes (e.g. collar, sleeve) and leaf landmark nodes (e.g. left-collar, right-coll
ar). Each Layout-Graph Reasoning(LGR) layer aims to map feature representations
into structural graph nodes via a Map-to-Node module, performs reasoning over st
ructural graph nodes to achieve global layout coherency via a layout-graph reaso
ning module, and then maps graph nodes back to enhance feature representations v
ia a Node-to-Map module. The layout-graph reasoning module integrates a graph cl
ustering operation to generate representations of intermediate nodes (bottom-up
inference) and then a graph deconvolution operation (top-down inference) over th
e whole graph. Extensive experiments on two public fashion landmark datasets dem
onstrate the superiority of our model. Furthermore, to advance the fine-grained
fashion landmark research for supporting more comprehensive clothes generation a
nd attribute recognition, we contribute the first Fine-grained Fashion Landmark
Dataset (FFLD) containing 200k images annotated with at most 32 key-points for 1
3 clothes types.
****

DistillHash: Unsupervised Deep Hashing by Distilling Data Pairs
Erkun Yang, Tongliang Liu, Cheng Deng, Wei Liu, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2946-2955
Due to storage and search efficiency, hashing has become significantly prevalent for nearest neighbor search. Particularly, deep hashing methods have greatly improved the search performance, typically under supervised scenarios. In contrast, unsupervised deep hashing models can hardly achieve satisfactory performance due to the lack of supervisory similarity signals. To address this problem, in this paper, we propose a new deep unsupervised hashing model, called DistilHash, which can learn a distilled data set, where data pairs have confident similarity signals. Specifically, we investigate the relationship between the initial but noisy similarity signals learned from local structures and the semantic similarity labels assigned by the optimal Bayesian classifier. We show that, under a mild assumption, some data pairs, of which labels are consistent with those assigned by the optimal Bayesian classifier, can be potentially distilled. With this understanding, we design a simple but effective method to distill data pairs automatically and further adopt a Bayesian learning framework to learn hashing functions from the distilled data set. Extensive experimental results on three widely used benchmark datasets demonstrate that our method achieves state-of-the-art search performance.
****
Mind Your Neighbours: Image Annotation With Metadata Neighbourhood Graph Co-Attention Networks
Junjie Zhang, Qi Wu, Jian Zhang, Chunhua Shen, Jianfeng Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2956-2964
As the visual reflections of our daily lives, images are frequently shared on the social network, which generates the abundant 'metadata' that records user interactions with images. Due to the diverse contents and complex styles, some images can be challenging to recognise when neglecting the context. Images with the similar metadata, such as 'relevant topics and textual descriptions', 'common friends of users' and 'nearby locations', form a neighbourhood for each image, which can be used to assist the annotation. In this paper, we propose a Metadata Neighbourhood Graph Co-Attention Network (MangoNet) to model the correlations between each target image and its neighbours. To accurately capture the visual clues from the neighbourhood, a co-attention mechanism is introduced to embed the target image and its neighbours as graph nodes, while the graph edges capture the node pair correlations. By reasoning on the neighbourhood graph, we obtain the graph representation to help annotate the target image. Experimental results on three benchmark datasets indicate that our proposed model achieves the best performance compared to the state-of-the-art methods.
****
Region Proposal by Guided Anchoring
Jiaqi Wang, Kai Chen, Shuo Yang, Chen Change Loy, Dahua Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2965-2974
Region anchors are the cornerstone of modern object detection techniques. State-of-the-art detectors mostly rely on a dense anchoring scheme, where anchors are sampled uniformly over the spatial domain with a predefined set of scales and aspect ratios. In this paper, we revisit this foundational stage. Our study shows that it can be done much more effectively and efficiently. Specifically, we present an alternative scheme, named Guided Anchoring, which leverages semantic features to guide the anchoring. The proposed method jointly predicts the locations where the center of objects of interest are likely to exist as well as the scales and aspect ratios at different locations. On top of predicted anchor shapes, we mitigate the feature inconsistency with a feature adaption module. We also study the use of high-quality proposals to improve detection performance. The anchoring scheme can be seamlessly integrated into proposal methods and detectors. With Guided Anchoring, we achieve 9.1% higher recall on MS COCO with 90% fewer anc

hors than the RPN baseline. We also adopt Guided Anchoring in Fast R-CNN, Faster R-CNN and RetinaNet, respectively improving the detection mAP by 2.2%, 2.7% and 1.2%. Code is available at https://github.com/open-mmlab/mmdetection.
****

Distant Supervised Centroid Shift: A Simple and Efficient Approach to Visual Domain Adaptation
Jian Liang, Ran He, Zhenan Sun, Tieniu Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2975-2984
Conventional domain adaptation methods usually resort to deep neural networks or subspace learning to find invariant representations across domains. However, most deep learning methods highly rely on large-size source domains and are computationally expensive to train, while subspace learning methods always have a quadratic time complexity that suffers from the large domain size. This paper provides a simple and efficient solution, which could be regarded as a well-performing baseline for domain adaptation tasks. Our method is built upon the nearest centroid classifier, seeking a subspace where the centroids in the target domain are moderately shifted from those in the source domain. Specifically, we design a unified objective without accessing the source domain data and adopt an alternating minimization scheme to iteratively discover the pseudo target labels, invariant subspace, and target centroids. Besides its privacy-preserving property (distant supervision), the algorithm is provably convergent and has a promising linear time complexity. In addition, the proposed method can be readily extended to multi-source setting and domain generalization, and it remarkably enhances popular deep adaptation methods by borrowing the learned transferable features. Extensive experiments on several benchmarks including object, digit, and face recognition datasets validate that our methods yield state-of-the-art results in various domain adaptation tasks.
****

Learning to Transfer Examples for Partial Domain Adaptation
Zhangjie Cao, Kaichao You, Mingsheng Long, Jianmin Wang, Qiang Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2985-2994
Domain adaptation is critical for learning in new and unseen environments. With domain adversarial training, deep networks can learn disentangled and transferable features that effectively diminish the dataset shift between the source and target domains for knowledge transfer. In the era of Big Data, large-scale labeled datasets are readily available, stimulating the interest in partial domain adaptation (PDA), which transfers a recognizer from a large labeled domain to a small unlabeled domain. It extends standard domain adaptation to the scenario where target labels are only a subset of source labels. Under the condition that target labels are unknown, the key challenges of PDA are how to transfer relevant examples in the shared classes to promote positive transfer and how to ignore irrelevant ones in the source domain to mitigate negative transfer. In this work, we propose a unified approach to PDA, Example Transfer Network (ETN), which jointly learns domain-invariant representations across domains and a progressive weighting scheme to quantify the transferability of source examples. A thorough evaluation on several benchmark datasets shows that ETN consistently achieves state-of-the-art results for various partial domain adaptation tasks.
****

Generalized Zero-Shot Recognition Based on Visually Semantic Embedding
Pengkai Zhu, Hanxiao Wang, Venkatesh Saligrama; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2995-3003
We propose a novel Generalized Zero-Shot learning (GZSL) method that is agnostic to both unseen images and unseen semantic vectors during training. Prior works in this context propose to map high-dimensional visual features to the semantic domain, which we believe contributes to the semantic gap. To bridge the gap, we propose a novel low-dimensional embedding of visual instances that is "visually semantic." Analogous to semantic data that quantifies the existence of an attribute in the presented instance, components of our visual embedding quantifies existence of a prototypical part-type in the presented instance. In parallel, as a

thought experiment, we quantify the impact of noisy semantic data by utilizing a novel visual oracle to visually supervise a learner. These factors, namely semantic noise, visual-semantic gap and label noise lead us to propose a new graphical model for inference with pairwise interactions between label, semantic data, and inputs. We tabulate results on a number of benchmark datasets demonstrating significant improvement in accuracy over state-of-art under both semantic and visual supervision.
****

Towards Visual Feature Translation
Jie Hu, Rongrong Ji, Hong Liu, Shengchuan Zhang, Cheng Deng, Qi Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3004-3013
Most existing visual search systems are deployed based upon fixed kinds of visual features, which prohibits the feature reusing across different systems or when upgrading systems with a new type of feature. Such a setting is obviously inflexible and time/memory consuming, which is indeed mendable if visual features can be "translated" across systems. In this paper, we make the first attempt towards visual feature translation to break through the barrier of using features across different visual search systems. To this end, we propose a Hybrid Auto-Encoder (HAE) to translate visual features, which learns a mapping by minimizing the translation and reconstruction errors. Based upon HAE, an Undirected Affinity Measurement (UAM) is further designed to quantify the affinity among different types of visual features. Extensive experiments have been conducted on several public datasets with sixteen different types of widely-used features in visual search systems. Quantitative results show the encouraging possibilities of feature translation. For the first time, the affinity among widely-used features like SIFT and DELF is reported.
****

Amodal Instance Segmentation With KINS Dataset
Lu Qi, Li Jiang, Shu Liu, Xiaoyong Shen, Jiaya Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3014-3023
Amodal instance segmentation, a new direction of instance segmentation, aims to segment each object instance involving its invisible, occluded parts to imitate human ability. This task requires to reason objects' complex structure. Despite important and futuristic, this task lacks data with large-scale and detailed annotations, due to the difficulty of correctly and consistently labeling invisible parts, which creates the huge barrier to explore the frontier of visual recognition. In this paper, we augment KITTI with more instance pixel-level annotation for 8 categories, which we call KITTI INStance dataset (KINS). We propose the network structure to reason invisible parts via a new multi-task framework with Multi-View Coding (MVC), which combines information in various recognition levels. Extensive experiments show that our MVC effectively improves both amodal and inmodal segmentation. The KINS dataset and our proposed method will be made publicly available.
****

Global Second-Order Pooling Convolutional Networks
Zilin Gao, Jiangtao Xie, Qilong Wang, Peihua Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3024-3033
Deep Convolutional Networks (ConvNets) are fundamental to, besides large-scale visual recognition, a lot of vision tasks. As the primary goal of the ConvNets is to characterize complex boundaries of thousands of classes in a high-dimensional space, it is critical to learn higher-order representations for enhancing non-linear modeling capability. Recently, Global Second-order Pooling (GSoP), plugged at the end of networks, has attracted increasing attentions, achieving much better performance than classical, first-order networks in a variety of vision tasks. However, how to effectively introduce higher-order representation in earlier layers for improving non-linear capability of ConvNets is still an open problem. In this paper, we propose a novel network model introducing GSoP across from l

ower to higher layers for exploiting holistic image information throughout a net work. Given an input 3D tensor outputted by some previous convolutional layer, w e perform GSoP to obtain a covariance matrix which, after nonlinear transformati on, is used for tensor scaling along channel dimension. Similarly, we can perfor m GSoP along spatial dimension for tensor scaling as well. In this way, we can m ake full use of the second-order statistics of the holistic image throughout a n etwork. The proposed networks are thoroughly evaluated on large-scale ImageNet-1 K, and experiments have shown that they outperform non-trivially the counterpart s while achieving state-of-the-art results.
****

Weakly Supervised Complementary Parts Models for Fine-Grained Image Classificati on From the Bottom Up
Weifeng Ge, Xiangru Lin, Yizhou Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3034-3043
Given a training dataset composed of images and corresponding category labels, d eep convolutional neural networks show a strong ability in mining discriminative parts for image classification. However, deep convolutional neural networks tra ined with image level labels only tend to focus on the most discriminative parts while missing other object parts, which could provide complementary information . In this paper, we approach this problem from a different perspective. We build complementary parts models in a weakly supervised manner to retrieve informatio n suppressed by dominant object parts detected by convolutional neural networks. Given image level labels only, we first extract rough object instances by perfo rming weakly supervised object detection and instance segmentation using Mask R-CNN and CRF-based segmentation. Then we estimate and search for the best parts m odel for each object instance under the principle of preserving as much diversit y as possible. In the last stage, we build a bi-directional long short-term memo ry (LSTM) network to fuze and encode the partial information of these complement ary parts into a comprehensive feature for image classification. Experimental re sults indicate that the proposed method not only achieves significant improvemen t over our baseline models, but also outperforms state-of-the-art algorithms by a large margin (6.7%, 2.8%, 5.2% respectively) on Stanford Dogs 120, Caltech-UCS D Birds 2011-200 and Caltech 256.
****

NetTailor: Tuning the Architecture, Not Just the Weights
Pedro Morgado, Nuno Vasconcelos; Proceedings of the IEEE/CVF Conference on Comp uter Vision and Pattern Recognition (CVPR), 2019, pp. 3044-3054
Real-world applications of object recognition often require the solution of mult iple tasks in a single platform. Under the standard paradigm of network fine-tun ing, an entirely new CNN is learned per task, and the final network size is inde pendent of task complexity. This is wasteful, since simple tasks require smaller networks than more complex tasks, and limits the number of tasks that can be so lved simultaneously. To address these problems, we propose a transfer learning p rocedure, denoted NetTailor, in which layers of a pre-trained CNN are used as un iversal blocks that can be combined with small task-specific layers to generate new networks. Besides minimizing classification error, the new network is traine d to mimic the internal activations of a strong unconstrained CNN, and minimize its complexity by the combination of 1) a soft-attention mechanism over blocks a nd 2) complexity regularization constraints. In this way, NetTailor can adapt th e network architecture, not just its weights, to the target task. Experiments sh ow that networks adapted to simple tasks, such as character or traffic sign reco gnition, become significantly smaller than those adapted to hard tasks, such as fine-grained recognition. More importantly, due to the modular nature of the pro cedure, this reduction in network complexity is achieved without compromise of e ither parameter sharing across tasks, or classification accuracy.
****

Learning-Based Sampling for Natural Image Matting
Jingwei Tang, Yagiz Aksoy, Cengiz Oztireli, Markus Gross, Tunc Ozan Aydin; P roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3055-3063

The goal of natural image matting is the estimation of opacities of a user-defined foreground object that is essential in creating realistic composite imagery. Natural matting is a challenging process due to the high number of unknowns in the mathematical modeling of the problem, namely the opacities as well as the foreground and background layer colors, while the original image serves as the single observation. In this paper, we propose the estimation of the layer colors through the use of deep neural networks prior to the opacity estimation. The layer color estimation is a better match for the capabilities of neural networks, and the availability of these colors substantially increase the performance of opacity estimation due to the reduced number of unknowns in the compositing equation. A prominent approach to matting in parallel to ours is called sampling-based matting, which involves gathering color samples from known-opacity regions to predict the layer colors. Our approach outperforms not only the previous hand-crafted sampling algorithms, but also current data-driven methods. We hence classify our method as a hybrid sampling- and learning-based approach to matting, and demonstrate the effectiveness of our approach through detailed ablation studies using alternative network architectures.

****

Learning Unsupervised Video Object Segmentation Through Visual Attention

Wenguan Wang, Hongmei Song, Shuyang Zhao, Jianbing Shen, Sanyuan Zhao, Steven C. H. Hoi, Haibin Ling; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3064-3074

This paper conducts a systematic study on the role of visual attention in Unsupervised Video Object Segmentation (UVOS) tasks. By elaborately annotating three popular video segmentation datasets (DAVIS, Youtube-Objects and SegTrack V2) with dynamic eye-tracking data in the UVOS setting, for the first time, we quantitatively verified the high consistency of visual attention behavior among human observers, and found strong correlation between human attention and explicit primary object judgements during dynamic, task-driven viewing. Such novel observations provide an in-depth insight into the underlying rationale behind UVOS. Inspired by these findings, we decouple UVOS into two sub-tasks: UVOS-driven Dynamic Visual Attention Prediction (DVAP) in spatiotemporal domain, and Attention-Guided Object Segmentation (AGOS) in spatial domain. Our UVOS solution enjoys three major merits: 1) modular training without using expensive video segmentation annotations, instead, using more affordable dynamic fixation data to train the initial video attention module and using existing fixation-segmentation paired static/image data to train the subsequent segmentation module; 2) comprehensive foreground understanding through multi-source learning; and 3) additional interpretability from the biologically-inspired and assessable attention. Experiments on popular benchmarks show that, even without using expensive video object mask annotations, our model achieves compelling performance in comparison with state-of-the-arts.

****

4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks

Christopher Choy, JunYoung Gwak, Silvio Savarese; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3075-3084

In many robotics and VR/AR applications, 3D-videos are readily-available input sources (a sequence of depth images, or LIDAR scans). However, in many cases, the 3D-videos are processed frame-by-frame either through 2D convnets or 3D perception algorithms. In this work, we propose 4-dimensional convolutional neural networks for spatio-temporal perception that can directly process such 3D-videos using high-dimensional convolutions. For this, we adopt sparse tensors and propose generalized sparse convolutions that encompass all discrete convolutions. To implement the generalized sparse convolution, we create an open-source auto-differentiation library for sparse tensors that provides extensive functions for high-dimensional convolutional neural networks. We create 4D spatio-temporal convolutional neural networks using the library and validate them on various 3D semantic segmentation benchmarks and proposed 4D datasets for 3D-video perception. To overcome challenges in 4D space, we propose the hybrid kernel, a special case of th

e generalized sparse convolution, and trilateral-stationary conditional random f
ields that enforce spatio-temporal consistency in the 7D space-time-chroma space
. Experimentally, we show that a convolutional neural network with only generali
zed 3D sparse convolutions can outperform 2D or 2D-3D hybrid methods by a large
margin. Also, we show that on 3D-videos, 4D spatio-temporal convolutional neural
 networks are robust to noise and outperform the 3D convolutional neural network
.
****
Pyramid Feature Attention Network for Saliency Detection
Ting Zhao, Xiangqian Wu; Proceedings of the IEEE/CVF Conference on Computer Vis
ion and Pattern Recognition (CVPR), 2019, pp. 3085-3094
Saliency detection is one of the basic challenges in computer vision. Recently,
CNNs are the most widely used and powerful techniques for saliency detection, in
 which feature maps from different layers are always integrated without distinct
ion. However, instinctively, the different feature maps of CNNs and the differen
t features in the same maps should play different roles in saliency detection. T
o address this problem, a novel CNN named pyramid feature attention network (PFA
N) is proposed to enhance the high-level context features and the low-level spat
ial structural features. In the proposed PFAN, a context-aware pyramid feature e
xtraction (CPFE) module is designed for multi-scale high-level feature maps to c
apture the rich context features. A channel-wise attention (CA) model and a spat
ial attention (SA) model are respectively applied to the CPFE feature maps and t
he low-level feature maps, and then fused to detect salient regions. Finally, an
 edge preservation loss is proposed to get the accurate boundaries of salient re
gions. The proposed PFAN is extensively evaluated on five benchmark datasets and
 the experimental results demonstrate that the proposed network outperforms the
state-of-the-art approaches under different evaluation metrics.
****
Co-Saliency Detection via Mask-Guided Fully Convolutional Networks With Multi-Sc
ale Label Smoothing
Kaihua Zhang, Tengpeng Li, Bo Liu, Qingshan Liu; Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3095-310
4
In image co-saliency detection problem, one critical issue is how to model the c
oncurrent pattern of the co-salient parts, which appears both within each image
and across all the relevant images. In this paper, we propose a hierarchical ima
ge co-saliency detection framework as a coarse to fine strategy to capture this
pattern. We first propose a mask-guided fully convolutional network structure to
 generate the initial co-saliency detection result. The mask is used for backgro
und removal and it is learned from the high-level feature response maps of the p
re-trained VGG-net output. We next propose a multi-scale label smoothing model t
o further refine the detection result. The proposed model jointly optimizes the
label smoothness of pixels and superpixels. Experiment results on three popular
image co-saliency detection benchmark datasets including iCoseg, MSRC and Cosal2
015 demonstrate the remarkable performance compared with the state-of-the-art me
thods.
****
SAIL-VOS: Semantic Amodal Instance Level Video Object Segmentation - A Synthetic
 Dataset and Baselines
Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang, Alexander G. Schwing
; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit
ion (CVPR), 2019, pp. 3105-3115
We introduce SAIL-VOS (Semantic Amodal Instance Level Video Object Segmentation)
, a new dataset aiming to stimulate semantic amodal segmentation research. Human
s can effortlessly recognize partially occluded objects and reliably estimate th
eir spatial extent beyond the visible. However, few modern computer vision techn
iques are capable of reasoning about occluded parts of an object. This is partly
 due to the fact that very few image datasets and no video dataset exist which p
ermit development of those methods. To address this issue, we present a syntheti
c dataset extracted from the photo-realistic game GTA-V. Each frame is accompani

ed with densely annotated, pixel-accurate visible and amodal segmentation masks with semantic labels. More than 1.8M objects are annotated resulting in 100 times more annotations than existing datasets. We demonstrate the challenges of the dataset by quantifying the performance of several baselines. Data and additional material is available at http://sailvos.web.illinois.edu.
****

Learning Instance Activation Maps for Weakly Supervised Instance Segmentation
Yi Zhu, Yanzhao Zhou, Huijuan Xu, Qixiang Ye, David Doermann, Jianbin Jiao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3116-3125
Discriminative region responses residing inside an object instance can be extracted from networks trained with image-level label supervision. However, learning the full extent of pixel-level instance response in a weakly supervised manner remains unexplored. In this work, we tackle this challenging problem by using a novel instance extent filling approach. We first design a process to selectively collect pseudo supervision from noisy segment proposals obtained with previously published techniques. The pseudo supervision is used to learn a differentiable filling module that predicts a class-agnostic activation map for each instance given the image and an incomplete region response. We refer to the above maps as Instance Activation Maps (IAMs), which provide a fine-grained instance-level representation and allow instance masks to be extracted by lightweight CRF. Extensive experiments on the PASCAL VOC12 dataset show that our approach beats the state-of-the-art weakly supervised instance segmentation methods by a significant margin and increases the inference speed by an order of magnitude. Our method also generalizes well across domains and to unseen object categories. Without fine-tuning for the specific tasks, our model trained on VOC12 dataset (20 classes) obtains top performance for weakly supervised object localization on the CUB dataset (200 classes) and achieves competitive results on three widely used salient object detection benchmarks.
****

Decoders Matter for Semantic Segmentation: Data-Dependent Decoding Enables Flexible Feature Aggregation
Zhi Tian, Tong He, Chunhua Shen, Youliang Yan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3126-3135
Recent semantic segmentation methods exploit encoder-decoder architectures to produce the desired pixel-wise segmentation prediction. The last layer of the decoders is typically a bilinear upsampling procedure to recover the final pixel-wise prediction. We empirically show that this oversimple and data-independent bilinear upsampling may lead to sub-optimal results. In this work, we propose a data-dependent upsampling (DUpsampling) to replace bilinear, which takes advantages of the redundancy in the label space of semantic segmentation and is able to recover the pixel-wise prediction from low-resolution outputs of CNNs. The main advantage of the new upsampling layer lies in that with a relatively lower-resolution feature map such as 1/16 or 1/32 of the input size, we can achieve even better segmentation accuracy, significantly reducing computation complexity. This is made possible by 1) the new upsampling layer's much improved reconstruction capability; and more importantly 2) the DUpsampling based decoder's flexibility in leveraging almost arbitrary combinations of the CNN encoders' features. Experiments on PASCAL VOC demonstrate that with much less computation complexity, our decoder outperforms the state-of-the-art decoder. Finally, without any post-processing, the framework equipped with our proposed decoder achieves new state-of-the-art performance on two datasets: 88.1% mIOU on PASCAL VOC with 30% computation of the previously best model; and 52.5% mIOU on PASCAL Context.
****

Box-Driven Class-Wise Region Masking and Filling Rate Guided Loss for Weakly Supervised Semantic Segmentation
Chunfeng Song, Yan Huang, Wanli Ouyang, Liang Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3136-3145
Semantic segmentation has achieved huge progress via adopting deep Fully Convolu

tional Networks (FCN). However, the performance of FCN based models severely rely on the amounts of pixel-level annotations which are expensive and time-consuming. To address this problem, it is a good choice to learn to segment with weak supervision from bounding boxes. How to make full use of the class-level and region-level supervisions from bounding boxes is the critical challenge for the weakly supervised learning task. In this paper, we first introduce a box-driven class-wise masking model (BCM) to remove irrelevant regions of each class. Moreover, based on the pixel-level segment proposal generated from the bounding box supervision, we could calculate the mean filling rates of each class to serve as an important prior cue, then we propose a filling rate guided adaptive loss (FR-Loss) to help the model ignore the wrongly labeled pixels in proposals. Unlike previous methods directly training models with the fixed individual segment proposals, our method can adjust the model learning with global statistical information. Thus it can help reduce the negative impacts from wrongly labeled proposals. We evaluate the proposed method on the challenging PASCAL VOC 2012 benchmark and compare with other methods. Extensive experimental results show that the proposed method is effective and achieves the state-of-the-art results.
****

Dual Attention Network for Scene Segmentation
Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, Hanqing Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3146-3154
In this paper, we address the scene segmentation task by capturing rich contextual dependencies based on the self-attention mechanism. Unlike previous works that capture contexts by multi-scale features fusion, we propose a Dual Attention Networks (DANet) to adaptively integrate local features with their global dependencies. Specifically, we append two types of attention modules on top of traditional dilated FCN, which model the semantic interdependencies in spatial and channel dimensions respectively. The position attention module selectively aggregates the features at each position by a weighted sum of the features at all positions. Similar features would be related to each other regardless of their distances. Meanwhile, the channel attention module selectively emphasizes interdependent channel maps by integrating associated features among all channel maps. We sum the outputs of the two attention modules to further improve feature representation which contributes to more precise segmentation results. We achieve new state-of-the-art segmentation performance on three challenging scene segmentation datasets, i.e., Cityscapes, PASCAL Context and COCO Stuff dataset. In particular, a Mean IoU score of 81.5% on Cityscapes test set is achieved without using coarse data.
****

InverseRenderNet: Learning Single Image Inverse Rendering
Ye Yu, William A. P. Smith; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3155-3164
We show how to train a fully convolutional neural network to perform inverse rendering from a single, uncontrolled image. The network takes an RGB image as input, regresses albedo and normal maps from which we compute lighting coefficients. Our network is trained using large uncontrolled image collections without ground truth. By incorporating a differentiable renderer, our network can learn from self-supervision. Since the problem is ill-posed we introduce additional supervision: 1. We learn a statistical natural illumination prior, 2. Our key insight is to perform offline multiview stereo (MVS) on images containing rich illumination variation. From the MVS pose and depth maps, we can cross project between overlapping views such that Siamese training can be used to ensure consistent estimation of photometric invariants. MVS depth also provides direct coarse supervision for normal map estimation. We believe this is the first attempt to use MVS supervision for learning inverse rendering.
****

A Variational Auto-Encoder Model for Stochastic Point Processes
Nazanin Mehrasa, Akash Abdu Jyothi, Thibaut Durand, Jiawei He, Leonid Sigal, Greg Mori; Proceedings of the IEEE/CVF Conference on Computer Vision and Patte

rn Recognition (CVPR), 2019, pp. 3165-3174
We propose a novel probabilistic generative model for action sequences. The model is termed the Action Point Process VAE (APP-VAE), a variational auto-encoder that can capture the distribution over the times and categories of action sequences. Modeling the variety of possible action sequences is a challenge, which we show can be addressed via the APP-VAE's use of latent representations and non-linear functions to parameterize distributions over which event is likely to occur next in a sequence and at what time. We empirically validate the efficacy of APP-VAE for modeling action sequences on the MultiTHUMOS and Breakfast datasets.
****

Unifying Heterogeneous Classifiers With Distillation
Jayakorn Vongkulbhisal, Phongtharin Vinayavekhin, Marco Visentini-Scarzanella; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3175-3184
In this paper, we study the problem of unifying knowledge from a set of classifiers with different architectures and target classes into a single classifier, given only a generic set of unlabelled data. We call this problem Unifying Heterogeneous Classifiers (UHC). This problem is motivated by scenarios where data is collected from multiple sources, but the sources cannot share their data, e.g., due to privacy concerns, and only privately trained models can be shared. In addition, each source may not be able to gather data to train all classes due to data availability at each source, and may not be able to train the same classification model due to different computational resources. To tackle this problem, we propose a generalisation of knowledge distillation to merge HCs. We derive a probabilistic relation between the outputs of HCs and the probability over all classes. Based on this relation, we propose two classes of methods based on cross-entropy minimisation and matrix factorisation, which allow us to estimate soft labels over all classes from unlabelled samples and use them in lieu of ground truth labels to train a unified classifier. Our extensive experiments on ImageNet, LSUN, and Places365 datasets show that our approaches significantly outperform a naive extension of distillation and can achieve almost the same accuracy as classifiers that are trained in a centralised, supervised manner.
****

Assessment of Faster R-CNN in Man-Machine Collaborative Search
Arturo Deza, Amit Surana, Miguel P. Eckstein; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3185-3194
With the advent of modern expert systems driven by deep learning that supplement human experts (e.g. radiologists, dermatologists, surveillance scanners), we analyze how and when do such expert systems enhance human performance in a fine-grained small target visual search task. We set up a 2 session factorial experimental design in which humans visually search for a target with and without a Deep Learning (DL) expert system. We evaluate human changes of target detection performance and eye-movements in the presence of the DL system. We find that performance improvements with the DL system (computed via a Faster R-CNN with a VGG16) interacts with observer's perceptual abilities (e.g., sensitivity). The main results include: 1) The DL system reduces the False Alarm rate per Image on average across observer groups of both high/low sensitivity; 2) Only human observers with high sensitivity perform better than the DL system, while the low sensitivity group does not surpass individual DL system performance, even when aided with the DL system itself; 3) Increases in number of trials and decrease in viewing time were mainly driven by the DL system only for the low sensitivity group. 4) The DL system aids the human observer to fixate at a target by the 3rd fixation. These results provide insights of the benefits and limitations of deep learning systems that are collaborative or competitive with humans.
****

OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge
Kenneth Marino, Mohammad Rastegari, Ali Farhadi, Roozbeh Mottaghi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3195-3204
Visual Question Answering (VQA) in its ideal form lets us study reasoning in the

joint space of vision and language and serves as a proxy for the AI task of sce
ne understanding. However, most VQA benchmarks to date are focused on questions
such as simple counting, visual attributes, and object detection that do not req
uire reasoning or knowledge beyond what is in the image. In this paper, we addre
ss the task of knowledge-based visual question answering and provide a benchmark
, called OK-VQA, where the image content is not sufficient to answer the questio
ns, encouraging methods that rely on external knowledge resources. Our new datas
et includes more than 14,000 questions that require external knowledge to answer
. We show that the performance of the state-of-the-art VQA models degrades drast
ically in this new setting. Our analysis shows that our knowledge-based VQA task
is diverse, difficult, and large compared to previous knowledge-based VQA datas
ets. We hope that this dataset enables researchers to open up new avenues for re
search in this domain.
****

NDDR-CNN: Layerwise Feature Fusing in Multi-Task CNNs by Neural Discriminative D
imensionality Reduction
Yuan Gao, Jiayi Ma, Mingbo Zhao, Wei Liu, Alan L. Yuille; Proceedings of the
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp
. 3205-3214
In this paper, we propose a novel Convolutional Neural Network (CNN) structure f
or general-purpose multi-task learning (MTL), which enables automatic feature fu
sing at every layer from different tasks. This is in contrast with the most wide
ly used MTL CNN structures which empirically or heuristically share features on
some specific layers (e.g., share all the features except the last convolutional
layer). The proposed layerwise feature fusing scheme is formulated by combining
existing CNN components in a novel way, with clear mathematical interpretabilit
y as discriminative dimensionality reduction, which is referred to as Neural Dis
criminative Dimensionality Reduction (NDDR). Specifically, we first concatenate
features with the same spatial resolution from different tasks according to thei
r channel dimension. Then, we show that the discriminative dimensionality reduct
ion can be fulfilled by 1x1 Convolution, Batch Normalization, and Weight Decay i
n one CNN. The use of existing CNN components ensures the end-to-end training an
d the extensibility of the proposed NDDR layer to various state-of-the-art CNN a
rchitectures in a "plug-and-play" manner. The detailed ablation analysis shows t
hat the proposed NDDR layer is easy to train and also robust to different hyperp
arameters. Experiments on different task sets with various base network architec
tures demonstrate the promising performance and desirable generalizability of ou
r proposed method. The code of our paper is available at https://github.com/etha
nygao/NDDR-CNN.
****

Spectral Metric for Dataset Complexity Assessment
Frederic Branchaud-Charron, Andrew Achkar, Pierre-Marc Jodoin; Proceedings of
the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019,
pp. 3215-3224
In this paper, we propose a new measure to gauge the complexity of image classif
ication problems. Given an annotated image dataset, our method computes a compl
exity measure called the cumulative spectral gradient (CSG) which strongly corre
lates with the test accuracy of convolutional neural networks (CNN). The CSG mea
sure is derived from the probabilistic divergence between classes in a spectral
clustering framework. We show that this metric correlates with the overall separ
ability of the dataset and thus its inherent complexity. As will be shown, our m
etric can be used for dataset reduction, to assess which classes are more diffic
ult to disentangle, and approximate the accuracy one could expect to get with a
CNN. Results obtained on 11 datasets and three CNN models reveal that our metho
d is more accurate and faster than previous complexity measures.
****

ADCrowdNet: An Attention-Injective Deformable Convolutional Network for Crowd Un
derstanding
Ning Liu, Yongchao Long, Changqing Zou, Qun Niu, Li Pan, Hefeng Wu; Proceed
ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR

), 2019, pp. 3225-3234
We propose an attention-injective deformable convolutional network called ADCrowdNet for crowd understanding that can address the accuracy degradation problem of highly congested noisy scenes. ADCrowdNet contains two concatenated networks. An attention-aware network called Attention Map Generator (AMG) first detects crowd regions in images and computes the congestion degree of these regions. Based on detected crowd regions and congestion priors, a multi-scale deformable network called Density Map Estimator (DME) then generates high-quality density maps. With the attention-aware training scheme and multi-scale deformable convolutional scheme, the proposed ADCrowdNet achieves the capability of being more effective to capture the crowd features and more resistant to various noises. We have evaluated our method on four popular crowd counting datasets (ShanghaiTech, UCF_CC_50, WorldEXPO'10, and UCSD) and an extra vehicle counting dataset TRANCOS, and our approach beats existing state-of-the-art approaches on all of these datasets.
****

VERI-Wild: A Large Dataset and a New Method for Vehicle Re-Identification in the Wild
Yihang Lou, Yan Bai, Jun Liu, Shiqi Wang, Lingyu Duan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3235-3243
Vehicle Re-identification (ReID) is of great significance to the intelligent transportation and public security. However, many challenging issues of Vehicle ReID in real-world scenarios have not been fully investigated, e.g., the high viewpoint variations, extreme illumination conditions, complex backgrounds, and different camera sources. To promote the research of vehicle ReID in the wild, we collect a new dataset called VERI-Wild with the following distinct features: 1) The vehicle images are captured by a large surveillance system containing 174 cameras covering a large urban district (more than 200km^2) The camera network continuously captures vehicles for 24 hours in each day and lasts for 1 month. 3) It is the first vehicle ReID dataset that is collected from unconstrained conditionsns. It is also a large dataset containing more than 400 thousand images of 40 thousand vehicle IDs. In this paper, we also propose a new method for vehicle ReID, in which, the ReID model is coupled into a Feature Distance Adversarial Network (FDA-Net), and a novel feature distance adversary scheme is designed to generate hard negative samples in feature space to facilitate ReID model training. The comprehensive results show the effectiveness of our method on the proposed dataset and the other two existing datasets.
****

3D Local Features for Direct Pairwise Registration
Haowen Deng, Tolga Birdal, Slobodan Ilic; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3244-3253
We present a novel, data driven approach for solving the problem of registration of two point cloud scans. Our approach is direct in the sense that a single pair of corresponding local patches already provides the necessary transformation cue for the global registration. To achieve that, we first endow the state of the art PPF-FoldNet auto-encoder (AE) with a pose-variant sibling, where the discrepancy between the two leads to pose-specific descriptors. Based upon this, we introduce RelativeNet, a relative pose estimation network to assign correspondence-specific orientations to the keypoints, eliminating any local reference frame computations. Finally, we devise a simple yet effective hypothesize-and-verify algorithm to quickly use the predictions and align two point sets. Our extensive quantitative and qualitative experiments suggests that our approach outperforms the state of the art in challenging real datasets of pairwise registration and that augmenting the keypoints with local pose information leads to better generalization and a dramatic speed-up.
****

HPLFlowNet: Hierarchical Permutohedral Lattice FlowNet for Scene Flow Estimation on Large-Scale Point Clouds
Xiuye Gu, Yijie Wang, Chongruo Wu, Yong Jae Lee, Panqu Wang; Proceedings of

the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3254-3263
We present a novel deep neural network architecture for end-to-end scene flow estimation that directly operates on large-scale 3D point clouds. Inspired by Bilateral Convolutional Layers (BCL), we propose novel DownBCL, UpBCL, and CorrBCL operations that restore structural information from unstructured point clouds, and fuse information from two consecutive point clouds. Operating on discrete and sparse permutohedral lattice points, our architectural design is parsimonious in computational cost. Our model can efficiently process a pair of point cloud frames at once with a maximum of 86K points per frame. Our approach achieves state-of-the-art performance on the FlyingThings3D and KITTI Scene Flow 2015 datasets. Moreover, trained on synthetic data, our approach shows great generalization ability on real-world data and on different point densities without fine-tuning.
****

GPSfM: Global Projective SFM Using Algebraic Constraints on Multi-View Fundamental Matrices
Yoni Kasten, Amnon Geifman, Meirav Galun, Ronen Basri; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3264-3272
This paper addresses the problem of recovering projective camera matrices from collections of fundamental matrices in multiview settings. We make two main contributions. First, given n \choose 2 fundamental matrices computed for n images, we provide a complete algebraic characterization in the form of conditions that are both necessary and sufficient to enabling the recovery of camera matrices. These conditions are based on arranging the fundamental matrices as blocks in a single matrix, called the n-view fundamental matrix, and characterizing this matrix in terms of the signs of its eigenvalues and rank structures. Secondly, we propose a concrete algorithm for projective structure-from-motion that utilizes this characterization. Given a complete or partial collection of measured fundamental matrices, our method seeks camera matrices that minimize a global algebraic error for the measured fundamental matrices. In contrast to existing methods, our optimization, without any initialization, produces a consistent set of fundamental matrices that corresponds to a unique set of cameras (up to a choice of projective frame). Our experiments indicate that our method achieves state of the art performance in both accuracy and running time.
****

Group-Wise Correlation Stereo Network
Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, Hongsheng Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3273-3282
Stereo matching estimates the disparity between a rectified image pair, which is of great importance to depth sensing, autonomous driving, and other related tasks. Previous works built cost volumes with cross-correlation or concatenation of left and right features across all disparity levels, and then a 2D or 3D convolutional neural network is utilized to regress the disparity maps. In this paper, we propose to construct the cost volume by group-wise correlation. The left features and the right features are divided into groups along the channel dimension, and correlation maps are computed among each group to obtain multiple matching cost proposals, which are then packed into a cost volume. Group-wise correlation provides efficient representations for measuring feature similarities and will not lose too much information like full correlation. It also preserves better performance when reducing parameters compared with previous methods. The 3D stacked hourglass network proposed in previous works is improved to boost the performance and decrease the inference computational cost. Experiment results show that our method outperforms previous methods on Scene Flow, KITTI 2012, and KITTI 2015 datasets.
****

Multi-Level Context Ultra-Aggregation for Stereo Matching
Guang-Yu Nie, Ming-Ming Cheng, Yun Liu, Zhengfa Liang, Deng-Ping Fan, Yue Liu, Yongtian Wang; Proceedings of the IEEE/CVF Conference on Computer Vision an

d Pattern Recognition (CVPR), 2019, pp. 3283-3291
Exploiting multi-level context information to cost volume can improve the performance of learning-based stereo matching methods. In recent years, 3-D Convolution Neural Networks (3-D CNNs) show the advantages in regularizing cost volume but are limited by unary features learning in matching cost computation. However, existing methods only use features from plain convolution layers or a simple aggregation of multi-level features to calculate cost volume, which is insufficient because stereo matching requires discriminative features to identify corresponding pixels in rectified stereo image pairs. In this paper, we propose a unary features descriptor using multi-level context ultra-aggregation (MCUA), which encapsulates all convolutional features into a more discriminative representation by intra- and inter-level features combination. Specifically, a child module that takes low-resolution images as input captures larger context information; the larger context information from each layer is densely connected to the main branch of the network. MCUA makes good usage of multi-level features with richer context and performs the image-to-image prediction holistically. We introduce our MCUA scheme for cost volume calculation and test it on PSM-Net. We also evaluate our method on Scene Flow and KITTI 2012/2015 stereo datasets. Experimental results show that our method outperforms state-of-the-art methods by a notable margin and effectively improves the accuracy of stereo matching.
****

Large-Scale, Metric Structure From Motion for Unordered Light Fields
Sotiris Nousias, Manolis Lourakis, Christos Bergeles; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3292-3301
This paper presents a large scale, metric Structure from Motion (SfM) pipeline for generalised cameras with overlapping fields-of-view, and demonstrates it using Light Field (LF) images. We build on recent developments in algorithms for absolute and relative pose recovery for generalised cameras and couple them with multi-view triangulation in a robust framework that advances the state-of-the-art on 3D reconstruction from LFs in several ways. First, our framework can recover the scale of a scene. Second, it is concerned with unordered sets of LF images, meticulously determining the order in which images should be considered. Third, it can scale to datasets with hundreds of LF images. Finally, it recovers 3D scene structure while abstaining from triangulating using very small baselines. Our approach outperforms the state-of-the-art, as demonstrated by real-world experiments with variable size datasets.
****

Understanding the Limitations of CNN-Based Absolute Camera Pose Regression
Torsten Sattler, Qunjie Zhou, Marc Pollefeys, Laura Leal-Taixe; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3302-3312
Visual localization is the task of accurate camera pose estimation in a known scene. It is a key problem in computer vision and robotics, with applications including self-driving cars, Structure-from-Motion, SLAM, and Mixed Reality. Traditionally, the localization problem has been tackled using 3D geometry. Recently, end-to-end approaches based on convolutional neural networks have become popular. These methods learn to directly regress the camera pose from an input image. However, they do not achieve the same level of pose accuracy as 3D structure-based methods. To understand this behavior, we develop a theoretical model for camera pose regression. We use our model to predict failure cases for pose regression techniques and verify our predictions through experiments. We furthermore use our model to show that pose regression is more closely related to pose approximation via image retrieval than to accurate pose estimation via 3D structure. A key result is that current approaches do not consistently outperform a handcrafted image retrieval baseline. This clearly shows that additional research is needed before pose regression algorithms are ready to compete with structure-based methods.
****

DeepLiDAR: Deep Surface Normal Guided Depth Prediction for Outdoor Scene From Sp

arse LiDAR Data and Single Color Image
Jiaxiong Qiu, Zhaopeng Cui, Yinda Zhang, Xingdi Zhang, Shuaicheng Liu, Bing Zeng, Marc Pollefeys; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3313-3322
In this paper, we propose a deep learning architecture that produces accurate dense depth for the outdoor scene from a single color image and a sparse depth. Inspired by the indoor depth completion, our network estimates surface normals as the intermediate representation to produce dense depth, and can be trained end-to-end. With a modified encoder-decoder structure, our network effectively fuses the dense color image and the sparse LiDAR depth. To address outdoor specific challenges, our network predicts a confidence mask to handle mixed LiDAR signals near foreground boundaries due to occlusion, and combines estimates from the color image and surface normals with learned attention maps to improve the depth accuracy especially for distant areas. Extensive experiments demonstrate that our model improves upon the state-of-the-art performance on KITTI depth completion benchmark. Ablation study shows the positive impact of each model components to the final performance, and comprehensive analysis shows that our model generalizes well to the input with higher sparsity or from indoor scenes.
****
Modeling Point Clouds With Self-Attention and Gumbel Subset Sampling
Jiancheng Yang, Qiang Zhang, Bingbing Ni, Linguo Li, Jinxian Liu, Mengdie Zhou, Qi Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3323-3332
Geometric deep learning is increasingly important thanks to the popularity of 3D sensors. Inspired by the recent advances in NLP domain, the self-attention transformer is introduced to consume the point clouds. We develop Point Attention Transformers (PATs), using a parameter-efficient Group Shuffle Attention (GSA) to replace the costly Multi-Head Attention. We demonstrate its ability to process size-varying inputs, and prove its permutation equivariance. Besides, prior work uses heuristics dependence on the input data (e.g., Furthest Point Sampling) to hierarchically select subsets of input points. Thereby, we for the first time propose an end-to-end learnable and task-agnostic sampling operation, named Gumbel Subset Sampling (GSS), to select a representative subset of input points. Equipped with Gumbel-Softmax, it produces a "soft" continuous subset in training phase, and a "hard" discrete subset in test phase. By selecting representative subsets in a hierarchical fashion, the networks learn a stronger representation of the input sets with lower computation cost. Experiments on classification and segmentation benchmarks show the effectiveness and efficiency of our methods. Furthermore, we propose a novel application, to process event camera stream as point clouds, and achieve a state-of-the-art performance on DVS128 Gesture Dataset.
****
Learning With Batch-Wise Optimal Transport Loss for 3D Shape Recognition
Lin Xu, Han Sun, Yuai Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3333-3342
Deep metric learning is essential for visual recognition. The widely used pair-wise (or triplet) based loss objectives cannot make full use of semantical information in training samples or give enough attention to those hard samples during optimization. Thus, they often suffer from a slow convergence rate and inferior performance. In this paper, we show how to learn an importance-driven distance metric via optimal transport programming from batches of samples. It can automatically emphasize hard examples and lead to significant improvements in convergence. We propose a new batch-wise optimal transport loss and combine it in an end-to-end deep metric learning manner. We use it to learn the distance metric and deep feature representation jointly for recognition. Empirical results on visual retrieval and classification tasks with six benchmark datasets, i.e., MNIST, CIFAR10, SHREC13, SHREC14, ModelNet10, and ModelNet40, demonstrate the superiority of the proposed method. It can accelerate the convergence rate significantly while achieving a state-of-the-art recognition performance. For example, in 3D shape recognition experiments, we show that our method can achieve better recognition performance within only 5 epochs than what can be obtained by mainstream 3

D shape recognition approaches after 200 epochs.
****

DenseFusion: 6D Object Pose Estimation by Iterative Dense Fusion
Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martin-Martin, Cewu Lu, Li Fei-Fei, Silvio Savarese; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3343-3352
A key technical challenge in performing 6D object pose estimation from RGB-D image is to fully leverage the two complementary data sources. Prior works either extract information from the RGB image and depth separately or use costly post-processing steps, limiting their performances in highly cluttered scenes and real-time applications. In this work, we present DenseFusion, a generic framework for estimating 6D pose of a set of known objects from RGB-D images. DenseFusion is a heterogeneous architecture that processes the two data sources individually and uses a novel dense fusion network to extract pixel-wise dense feature embedding, from which the pose is estimated. Furthermore, we integrate an end-to-end iterative pose refinement procedure that further improves the pose estimation while achieving near real-time inference. Our experiments show that our method outperforms state-of-the-art approaches in two datasets, YCB-Video and LineMOD. We also deploy our proposed method to a real robot to grasp and manipulate objects based on the estimated pose.
****

Dense Depth Posterior (DDP) From Single Image and Sparse Range
Yanchao Yang, Alex Wong, Stefano Soatto; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3353-3362
We present a deep learning system to infer the posterior distribution of a dense depth map associated with an image, by exploiting sparse range measurements, for instance from a lidar. While the lidar may provide a depth value for a small percentage of the pixels, we exploit regularities reflected in the training set to complete the map so as to have a probability over depth for each pixel in the image. We exploit a Conditional Prior Network, that allows associating a probability to each depth value given an image, and combine it with a likelihood term that uses the sparse measurements. Optionally we can also exploit the availability of stereo during training, but in any case only require a single image and a sparse point cloud at run-time. We test our approach on both unsupervised and supervised depth completion using the KITTI benchmark, and improve the state-of-the-art in both.
****

DuLa-Net: A Dual-Projection Network for Estimating Room Layouts From a Single RGB Panorama
Shang-Ta Yang, Fu-En Wang, Chi-Han Peng, Peter Wonka, Min Sun, Hung-Kuo Chu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3363-3372
We present a deep learning framework, called DuLa-Net, to predict Manhattan-world 3D room layouts from a single RGB panorama. To achieve better prediction accuracy, our method leverages two projections of the panorama at once, namely the equirectangular panorama-view and the perspective ceiling-view, that each contains different clues about the room layouts. Our network architecture consists of two encoder-decoder branches for analyzing each of the two views. In addition, a novel feature fusion structure is proposed to connect the two branches, which are then jointly trained to predict the 2D floor plans and layout heights. To learn more complex room layouts, we introduce the Realtor360 dataset that contains panoramas of Manhattan-world room layouts with different numbers of corners. Experimental results show that our work outperforms recent state-of-the-art in prediction accuracy and performance, especially in the rooms with non-cuboid layouts.
****

Veritatem Dies Aperit - Temporally Consistent Depth Prediction Enabled by a Multi-Task Geometric and Semantic Scene Understanding Approach
Amir Atapour-Abarghouei, Toby P. Breckon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3373-3384

Robust geometric and semantic scene understanding is ever more important in many real-world applications such as autonomous driving and robotic navigation. In this paper, we propose a multi-task learning-based approach capable of jointly performing geometric and semantic scene understanding, namely depth prediction (monocular depth estimation and depth completion) and semantic scene segmentation. Within a single temporally constrained recurrent network, our approach uniquely takes advantage of a complex series of skip connections, adversarial training and the temporal constraint of sequential frame recurrence to produce consistent depth and semantic class labels simultaneously. Extensive experimental evaluation demonstrates the efficacy of our approach compared to other contemporary state-of-the-art techniques.
****

Segmentation-Driven 6D Object Pose Estimation
Yinlin Hu, Joachim Hugonot, Pascal Fua, Mathieu Salzmann; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3385-3394
The most recent trend in estimating the 6D pose of rigid objects has been to train deep networks to either directly regress the pose from the image or to predict the 2D locations of 3D keypoints, from which the pose can be obtained using a PnP algorithm. In both cases, the object is treated as a global entity, and a single pose estimate is computed. As a consequence, the resulting techniques can be vulnerable to large occlusions. In this paper, we introduce a segmentation-driven 6D pose estimation framework where each visible part of the objects contributes a local pose prediction in the form of 2D keypoint locations. We then use a predicted measure of confidence to combine these pose candidates into a robust set of 3D-to-2D correspondences, from which a reliable pose estimate can be obtained. We outperform the state-of-the-art on the challenging Occluded-LINEMOD and YCB-Video datasets, which is evidence that our approach deals well with multiple poorly-textured objects occluding each other. Furthermore, it relies on a simple enough architecture to achieve real-time performance.
****

Exploiting Temporal Context for 3D Human Pose Estimation in the Wild
Anurag Arnab, Carl Doersch, Andrew Zisserman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3395-3404
We present a bundle-adjustment-based algorithm for recovering accurate 3D human pose and meshes from monocular videos. Unlike previous algorithms which operate on single frames, we show that reconstructing a person over an entire sequence gives extra constraints that can resolve ambiguities. This is because videos often give multiple views of a person, yet the overall body shape does not change and 3D positions vary slowly. Our method improves not only on standard mocap-based datasets like Human 3.6M -- where we show quantitative improvements -- but also on challenging in-the-wild datasets such as Kinetics. Building upon our algorithm, we present a new dataset of more than 3 million frames of YouTube videos from Kinetics with automatically generated 3D poses and meshes. We show that retraining a single-frame 3D pose estimator on this data improves accuracy on both real-world and mocap data by evaluating on the 3DPW and HumanEVA datasets.
****

What Do Single-View 3D Reconstruction Networks Learn?
Maxim Tatarchenko, Stephan R. Richter, Rene Ranftl, Zhuwen Li, Vladlen Koltun, Thomas Brox; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3405-3414
Convolutional networks for single-view object reconstruction have shown impressive performance and have become a popular subject of research. All existing techniques are united by the idea of having an encoder-decoder network that performs non-trivial reasoning about the 3D structure of the output space. In this work, we set up two alternative approaches that perform image classification and retrieval respectively. These simple baselines yield better results than state-of-the-art methods, both qualitatively and quantitatively. We show that encoder-decoder methods are statistically indistinguishable from these baselines, thus indicating that the current state of the art in single-view object reconstruction does

not actually perform reconstruction but image classification. We identify aspects of popular experimental procedures that elicit this behavior and discuss ways to improve the current state of research.
****

UniformFace: Learning Deep Equidistributed Representation for Face Recognition
Yueqi Duan, Jiwen Lu, Jie Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3415-3424
In this paper, we propose a new supervision objective named uniform loss to learn deep equidistributed representations for face recognition. Most existing methods aim to learn discriminative face features, encouraging large inter-class distances and small intra-class variations. However, they ignore the distribution of faces in the holistic feature space, which may lead to severe locality and unbalance. With the prior that faces lie on a hypersphere manifold, we impose an equidistributed constraint by uniformly spreading the class centers on the manifold, so that the minimum distance between class centers can be maximized through complete exploitation of the feature space. To this end, we consider the class centers as like charges on the surface of hypersphere with inter-class repulsion, and minimize the total electric potential energy as the uniform loss. Extensive experimental results on the MegaFace Challenge I, IARPA Janus Benchmark A (IJB-A), Youtube Faces (YTF) and Labeled Faces in the Wild (LFW) datasets show the effectiveness of the proposed uniform loss.
****

Semantic Graph Convolutional Networks for 3D Human Pose Regression
Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, Dimitris N. Metaxas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3425-3435
In this paper, we study the problem of learning Graph Convolutional Networks (GCNs) for regression. Current architectures of GCNs are limited to the small receptive field of convolution filters and shared transformation matrix for each node. To address these limitations, we propose Semantic Graph Convolutional Networks (SemGCN), a novel neural network architecture that operates on regression tasks with graph-structured data. SemGCN learns to capture semantic information such as local and global node relationships, which is not explicitly represented in the graph. These semantic relationships can be learned through end-to-end training from the ground truth without additional supervision or hand-crafted rules. We further investigate applying SemGCN to 3D human pose regression. Our formulation is intuitive and sufficient since both 2D and 3D human poses can be represented as a structured graph encoding the relationships between joints in the skeleton of a human body. We carry out comprehensive studies to validate our method. The results prove that SemGCN outperforms state of the art while using 90% fewer parameters.
****

Mask-Guided Portrait Editing With Conditional GANs
Shuyang Gu, Jianmin Bao, Hao Yang, Dong Chen, Fang Wen, Lu Yuan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3436-3445
Portrait editing is a popular subject in photo manipulation.The Generative Adversarial Network (GAN) advances the generating of realistic faces and allows more face editing. In this paper, we argue about three issues in existing techniques: diversity, quality, and controllability for portrait synthesis and editing. To address these issues, we propose a novel end-to-end learning framework that leverages conditional GANs guided by provided face masks for generating faces. The framework learns feature embeddings for every face component (e.g., mouth, hair, eye), separately, contributing to better correspondences for image translation, and local face editing. With the mask, our network is available to many applications, like face synthesis driven by mask, face Swap+ (including hair in swapping), and local manipulation. It can also boost the performance of face parsing a bit as an option of data augmentation.
****

Group Sampling for Scale Invariant Face Detection

Xiang Ming, Fangyun Wei, Ting Zhang, Dong Chen, Fang Wen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3446-3456

Detectors based on deep learning tend to detect multi-scale faces on a single input image for efficiency. Recent works, such as FPN and SSD, generally use feature maps from multiple layers with different spatial resolutions to detect objects at different scales, e.g., high-resolution feature maps for small objects. However, we find that such multi-layer prediction is not necessary. Faces at all scales can be well detected with features from a single layer of the network. In this paper, we carefully examine the factors affecting face detection across a large range of scales, and conclude that the balance of training samples, including both positive and negative ones, at different scales is the key. We propose a group sampling method which divides the anchors into several groups according to the scale, and ensure that the number of samples for each group is the same during training. Our approach using only the last layer of FPN as features is able to advance the state-of-the-arts. Comprehensive analysis and extensive experiments have been conducted to show the effectiveness of the proposed method. Our approach, evaluated on face detection benchmarks including FDDB and WIDER FACE datasets, achieves state-of-the-art results without bells and whistles.

****

Joint Representation and Estimator Learning for Facial Action Unit Intensity Estimation

Yong Zhang, Baoyuan Wu, Weiming Dong, Zhifeng Li, Wei Liu, Bao-Gang Hu, Qiang Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3457-3466

Facial action unit (AU) intensity is an index to characterize human expressions. Accurate AU intensity estimation depends on three major elements: image representation, intensity estimator, and supervisory information. Most existing methods learn intensity estimator with fixed image representation, and rely on the availability of fully annotated supervisory information. In this paper, a novel general framework for AU intensity estimation is presented, which differs from traditional estimation methods in two aspects. First, rather than keeping image representation fixed, it simultaneously learns representation and intensity estimator to achieve an optimal solution. Second, it allows incorporating weak supervisory training signal from human knowledge (e.g. feature smoothness, label smoothness, label ranking, and positive label), which makes our model trainable even fully annotated information is not available. More specifically, human knowledge is represented as either soft or hard constraints which are encoded as regularization terms or equality/inequality constraints, respectively. On top of our novel framework, we additionally propose an efficient algorithm for optimization based on Alternating Direction Method of Multipliers (ADMM). Evaluations on two benchmark databases show that our method outperforms competing methods under different ratios of AU intensity annotations, especially for small ratios.

****

Semantic Alignment: Finding Semantically Consistent Ground-Truth for Facial Landmark Detection

Zhiwei Liu, Xiangyu Zhu, Guosheng Hu, Haiyun Guo, Ming Tang, Zhen Lei, Neil M. Robertson, Jinqiao Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3467-3476

Recently, deep learning based facial landmark detection has achieved great success. Despite this, we notice that the semantic ambiguity greatly degrades the detection performance. Specifically, the semantic ambiguity means that some landmarks (e.g. those evenly distributed along the face contour) do not have clear and accurate definition, causing the inconsistent annotations (random errors) introduced by annotators. Accordingly, these inconsistent annotations, which are usually provided by public databases, commonly work as the (inaccurate) groundtruth to supervise network training, leading to the degraded accuracy. To our knowledge, very little research has investigated this problem. In this paper, we propose a novel probabilistic model which introduces a latent variable, i.e. 'real' groundtruth which is semantically consistent, to optimize. This framework couples tw

o parts (1) training landmark detection CNN and (2) searching the 'real' groundt ruth. These two parts are alternatively optimized: the searched 'real' groundtru th supervises the CNN training; and the trained CNN assists the searching of 're al' groundtruth. In addition, to correct or recover the unconfidently predicted landmarks due to occlusion and low quality, we propose a global heatmap correcti on unit (GHCU) to correct outliers by considering the global face shape as a con straint. Extensive experiments on both image-based (300V and AFLW) and video-bas ed (300VW) databases demonstrate that our method effectively improves the landma rk detection accuracy and achieves state-of-the-art performance.
****

LAEO-Net: Revisiting People Looking at Each Other in Videos
Manuel J. Marin-Jimenez, Vicky Kalogeiton, Pablo Medina-Suarez, Andrew Zisser man; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog nition (CVPR), 2019, pp. 3477-3485
Capturing the 'mutual gaze' of people is essential for understanding and interpr eting the social interactions between them. To this end, this paper addresses th e problem of detecting people Looking At Each Other (LAEO) in video sequences. F or this purpose, we propose LAEO-Net, a new deep CNN for determining LAEO in vid eos. In contrast to previous works, LAEO-Net takes spatio-temporal tracks as inp ut and reasons about the whole track. It consists of three branches, one for eac h character's tracked head and one for their relative position. Moreover, we int roduce two new LAEO datasets: UCO-LAEO and AVA-LAEO. A thorough experimental eva luation demonstrates the ability of LAEO-Net to successfully determine if two pe ople are LAEO and the temporal window where it happens. Our model achieves state -of-the-art results on the existing TVHID-LAEO video dataset, significantly outp erforming previous approaches.
****

Robust Facial Landmark Detection via Occlusion-Adaptive Deep Networks
Meilu Zhu, Daming Shi, Mingjie Zheng, Muhammad Sadiq; Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 348 6-3496
In this paper, we present a simple and effective framework called Occlusion-adap tive Deep Networks (ODN) with the purpose of solving the occlusion problem for f acial landmark detection. In this model, the occlusion probability of each posit ion in high-level features are inferred by a distillation module that can be lea rnt automatically in the process of estimating the relationship between facial a ppearance and facial shape. The occlusion probability serves as the adaptive wei ght on high-level features to reduce the impact of occlusion and obtain clean fe ature representation. Nevertheless, the clean feature representation cannot repr esent the holistic face due to the missing semantic features. To obtain exhausti ve and complete feature representation, it is vital that we leverage a low-rank learning module to recover lost features. Considering that facial geometric char acteristics are conducive to the low-rank module to recover lost features, we pr opose a geometry-aware module to excavate geometric relationships between differ ent facial components. Depending on the synergistic effect of three modules, the proposed network achieves better performance in comparison to state-of-the-art methods on challenging benchmark datasets.
****

Learning Individual Styles of Conversational Gesture
Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, Jitendr a Malik; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern R ecognition (CVPR), 2019, pp. 3497-3506
Human speech is often accompanied by hand and arm gestures. We present a method for cross-modal translation from "in-the-wild" monologue speech of a single spea ker to their conversational gesture motion. We train on unlabeled videos for whi ch we only have noisy pseudo ground truth from an automatic pose detection syste m. Our proposed model significantly outperforms baseline methods in a quantitati ve comparison. To support research toward obtaining a computational understandin g of the relationship between gesture and speech, we release a large video datas et of person-specific gestures.

****
Face Anti-Spoofing: Model Matters, so Does Data
Xiao Yang, Wenhan Luo, Linchao Bao, Yuan Gao, Dihong Gong, Shibao Zheng, Zhifeng Li, Wei Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3507-3516
Face anti-spoofing is an important task in full-stack face applications including face detection, verification, and recognition. Previous approaches build models on datasets which do not simulate the real-world data well (e.g., small scale, insignificant variance, etc.). Existing models may rely on auxiliary information, which prevents these anti-spoofing solutions from generalizing well in practice. In this paper, we present a data collection solution along with a data synthesis technique to simulate digital medium-based face spoofing attacks, which can easily help us obtain a large amount of training data well reflecting the real-world scenarios. Through exploiting a novel Spatio-Temporal Anti-Spoof Network (STASN), we are able to push the performance on public face anti-spoofing datasets over state-of-the-art methods by a large margin. Since the proposed model can automatically attend to discriminative regions, it makes analyzing the behaviors of the network possible.We conduct extensive experiments and show that the proposed model can distinguish spoof faces by extracting features from a variety of regions to seek out subtle evidences such as borders, moire patterns, reflection artifacts, etc.
****
Fast Human Pose Estimation
Feng Zhang, Xiatian Zhu, Mao Ye; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3517-3526
Existing human pose estimation approaches often only consider how to improve the model generalisation performance, but putting aside the significant efficiency problem. This leads to the development of heavy models with poor scalability and cost-effectiveness in practical use. In this work, we investigate the under-studied but practically critical pose model efficiency problem. To this end, we present a new Fast Pose Distillation (FPD) model learning strategy. Specifically, the FPD trains a lightweight pose neural network architecture capable of executing rapidly with low computational cost. It is achieved by effectively transferring the pose structure knowledge of a strong teacher network. Extensive evaluations demonstrate the advantages of our FPD method over a broad range of state-of-the-art pose estimation approaches in terms of model cost-effectiveness on two standard benchmark datasets, MPII Human Pose and Leeds Sports Pose.
****
Decorrelated Adversarial Learning for Age-Invariant Face Recognition
Hao Wang, Dihong Gong, Zhifeng Li, Wei Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3527-3536
There has been an increasing research interest in age-invariant face recognition. However, matching faces with big age gaps remains a challenging problem, primarily due to the significant discrepancy of face appearance caused by aging. To reduce such discrepancy, in this paper we present a novel algorithm to remove age-related components from features mixed with both identity and age information. Specifically, we factorize a mixed face feature into two uncorrelated components: identity-dependent component and age-dependent component, where the identity-dependent component contains information that is useful for face recognition. To implement this idea, we propose the Decorrelated Adversarial Learning (DAL) algorithm, where a Canonical Mapping Module (CMM) is introduced to find maximum correlation of the paired features generated by the backbone network, while the backbone network and the factorization module are trained to generate features reducing the correlation. Thus, the proposed model learns the decomposed features of age and identity whose correlation is significantly reduced. Simultaneously, the identity-dependent feature and the age-dependent feature are supervised by ID and age preserving signals respectively to ensure they contain the correct information. Extensive experiments have been conducted on the popular public-domain face aging datasets (FG-NET, MORPH Album 2, and CACD-VS) to demonstrate the effectiveness of the proposed approach.

****
Cross-Task Weakly Supervised Learning From Instructional Videos
Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, Josef Sivic; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3537-3545
In this paper we investigate learning visual models for the steps of ordinary tasks using weak supervision via instructional narrations and an ordered list of steps instead of strong supervision via temporal annotations. At the heart of our approach is the observation that weakly supervised learning may be easier if a model shares components while learning different steps: "pour egg" should be trained jointly with other tasks involving "pour" and "egg". We formalize this in a component model for recognizing steps and a weakly supervised learning framework that can learn this model under temporal constraints from narration and the list of steps. Past data does not permit systematic studying of sharing and so we also gather a new dataset aimed at assessing cross-task sharing. Our experiments demonstrate that sharing across tasks improves performance, especially when done at the component level and that our component model can parse previously unseen tasks by virtue of its compositionality.
****
D3TW: Discriminative Differentiable Dynamic Time Warping for Weakly Supervised Action Alignment and Segmentation
Chien-Yi Chang, De-An Huang, Yanan Sui, Li Fei-Fei, Juan Carlos Niebles; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3546-3555
We address weakly supervised action alignment and segmentation in videos, where only the order of occurring actions is available during training. We propose Discriminative Differentiable Dynamic Time Warping (D3TW), the first discriminative model using weak ordering supervision. The key technical challenge for discriminative modeling with weak supervision is that the loss function of the ordering supervision is usually formulated using dynamic programming and is thus not differentiable. We address this challenge with a continuous relaxation of the min-operator in dynamic programming and extend the alignment loss to be differentiable. The proposed D3TW innovatively solves sequence alignment with discriminative modeling and end-to-end training, which substantially improves the performance in weakly supervised action alignment and segmentation tasks. We show that our model is able to bypass the degenerated sequence problem usually encountered in previous work and outperform the current state-of-the-art across three evaluation metrics in two challenging datasets.
****
Progressive Teacher-Student Learning for Early Action Prediction
Xionghui Wang, Jian-Fang Hu, Jian-Huang Lai, Jianguo Zhang, Wei-Shi Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3556-3565
The goal of early action prediction is to recognize actions from partially observed videos with incomplete action executions, which is quite different from action recognition. Predicting early actions is very challenging since the partially observed videos do not contain enough action information for recognition. In this paper, we aim at improving early action prediction by proposing a novel teacher-student learning framework. Our framework involves a teacher model for recognizing actions from full videos, a student model for predicting early actions from partial videos, and a teacher-student learning block for distilling progressive knowledge from teacher to student, crossing different tasks. Extensive experiments on three public action datasets show that the proposed progressive teacher-student learning framework can consistently improve performance of early action prediction model. We have also reported the state-of-the-art performances for early action prediction on all of these sets.
****
Social Relation Recognition From Videos via Multi-Scale Spatial-Temporal Reasoning
Xinchen Liu, Wu Liu, Meng Zhang, Jingwen Chen, Lianli Gao, Chenggang Yan,

Tao Mei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3566-3574
Discovering social relations, e.g., kinship, friendship, etc., from visual contents can make machines better interpret the behaviors and emotions of human beings. Existing studies mainly focus on recognizing social relations from still images while neglecting another important media--video. On one hand, the actions and storylines in videos provide more important cues for social relation recognition. On the other hand, the key persons may appear at arbitrary spatial-temporal locations, even not in one same image from beginning to the end. To overcome these challenges, we propose a Multi-scale Spatial-Temporal Reasoning (MSTR) framework to recognize social relations from videos. For the spatial representation, we not only adopt a temporal segment network to learn global action and scene information, but also design a Triple Graphs model to capture visual relations between persons and objects. For the temporal domain, we propose a Pyramid Graph Convolutional Network to perform temporal reasoning with multi-scale receptive fields, which can obtain both long-term and short-term storylines in videos. By this means, MSTR can comprehensively explore the multi-scale actions and storylines in spatial-temporal dimensions for social relation reasoning in videos. Extensive experiments on a new large-scale Video Social Relation dataset demonstrate the effectiveness of the proposed framework.
****

MS-TCN: Multi-Stage Temporal Convolutional Network for Action Segmentation
Yazan Abu Farha, Jurgen Gall; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3575-3584
Temporally locating and classifying action segments in long untrimmed videos is of particular interest to many applications like surveillance and robotics. While traditional approaches follow a two-step pipeline, by generating frame-wise probabilities and then feeding them to high-level temporal models, recent approaches use temporal convolutions to directly classify the video frames. In this paper, we introduce a multi-stage architecture for the temporal action segmentation task. Each stage features a set of dilated temporal convolutions to generate an initial prediction that is refined by the next one. This architecture is trained using a combination of a classification loss and a proposed smoothing loss that penalizes over-segmentation errors. Extensive evaluation shows the effectiveness of the proposed model in capturing long-range dependencies and recognizing action segments. Our model achieves state-of-the-art results on three challenging datasets: 50Salads, Georgia Tech Egocentric Activities (GTEA), and the Breakfast dataset.
****

Transferable Interactiveness Knowledge for Human-Object Interaction Detection
Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yanfeng Wang, Cewu Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3585-3594
Human-Object Interaction (HOI) Detection is an important problem to understand how humans interact with objects. In this paper, we explore Interactiveness Knowledge which indicates whether human and object interact with each other or not. We found that interactiveness knowledge can be learned across HOI datasets, regardless of HOI category settings. Our core idea is to exploit an Interactiveness Network to learn the general interactiveness knowledge from multiple HOI datasets and perform Non-Interaction Suppression before HOI classification in inference. On account of the generalization of interactiveness, interactiveness network is a transferable knowledge learner and can be cooperated with any HOI detection models to achieve desirable results. We extensively evaluate the proposed method on HICO-DET and V-COCO datasets. Our framework outperforms state-of-the-art HOI detection results by a great margin, verifying its efficacy and flexibility. Code is available at https://github.com/DirtyHarryLYL/Transferable-Interactiveness-Network.
****

Actional-Structural Graph Convolutional Networks for Skeleton-Based Action Recognition

Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, Qi Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3595-3603

Action recognition with skeleton data has recently attracted much attention in computer vision. Previous studies are mostly based on fixed skeleton graphs, only capturing local physical dependencies among joints, which may miss implicit joint correlations. To capture richer dependencies, we introduce an encoder-decoder structure, called A-link inference module, to capture action-specific latent dependencies, i.e. actional links, directly from actions. We also extend the existing skeleton graphs to represent higher-order dependencies, i.e. structural links. Combing the two types of links into a generalized skeleton graph, We further propose the actional-structural graph convolution network (AS-GCN), which stacks actional-structural graph convolution and temporal convolution as a basic building block, to learn both spatial and temporal features for action recognition. A future pose prediction head is added in parallel to the recognition head to help capture more detailed action patterns through self-supervision. We validate AS-GCN in action recognition using two skeleton data sets, NTU-RGB+D and Kinetics. The proposed AS-GCN achieves consistently large improvement compared to the state-of-the-art methods. As a side product, AS-GCN also shows promising results for future pose prediction.

****

Multi-Granularity Generator for Temporal Action Proposal

Yuan Liu, Lin Ma, Yifeng Zhang, Wei Liu, Shih-Fu Chang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3604-3613

Temporal action proposal generation is an important task, aiming to localize the video segments containing human actions in an untrimmed video. In this paper, we propose a multi-granularity generator (MGG) to perform the temporal action proposal from different granularity perspectives, relying on the video visual features equipped with the position embedding information. First, we propose to use a bilinear matching model to exploit the rich local information within the video sequence. Afterwards, two components, namely segment proposal producer (SPP) and frame actionness producer (FAP), are combined to perform the task of temporal action proposal at two distinct granularities. SPP considers the whole video in the form of feature pyramid and generates segment proposals from one coarse perspective, while FAP carries out a finer actionness evaluation for each video frame. Our proposed MGG can be trained in an end-to-end fashion. Through temporally adjusting the segment proposals with fine-grained information based on frame actionness, MGG achieves the superior performance over state-of-the-art methods on the public THUMOS-14 and ActivityNet-1.3 datasets. Moreover, we employ existing action classifiers to perform the classification of the proposals generated by MGG, leading to significant improvements compared against the competing methods for the video detection task.

****

Deep Rigid Instance Scene Flow

Wei-Chiu Ma, Shenlong Wang, Rui Hu, Yuwen Xiong, Raquel Urtasun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3614-3622

In this paper we tackle the problem of scene flow estimation in the context of self-driving. We leverage deep learning techniques as well as strong priors as in our application domain the motion of the scene can be composed by the motion of the robot and the 3D motion of the actors in the scene. We formulate the problem as energy minimization in a deep structured model, which can be solved efficiently in the GPU by unrolling a Gaussian-Newton solver. Our experiments in the challenging KITTI scene flow dataset show that we outperform the state-of-the-art by a very large margin, while being 800 times faster.

****

See More, Know More: Unsupervised Video Object Segmentation With Co-Attention Siamese Networks

Xiankai Lu, Wenguan Wang, Chao Ma, Jianbing Shen, Ling Shao, Fatih Porikli;

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti
on (CVPR), 2019, pp. 3623-3632
We introduce a novel network, called as CO-attention Siamese Network (COSNet), t
o address the unsupervised video object segmentation task from a holistic view.
We emphasize the importance of inherent correlation among video frames and incor
porate a global co-attention mechanism to improve further the state-of-the-art d
eep learning based solutions that primarily focus on learning discriminative for
eground representations over appearance and motion in short-term temporal segmen
ts. The co-attention layers in our network provide efficient and competent stage
s for capturing global correlations and scene context by jointly computing and a
ppending co-attention responses into a joint feature space. We train COSNet with
pairs of video frames, which naturally augments training data and allows increa
sed learning capacity. During the segmentation stage, the co-attention model enc
odes useful information by processing multiple reference frames together, which
is leveraged to infer the frequently reappearing and salient foreground objects
better. We propose a unified and end-to-end trainable framework where different
co-attention variants can be derived for mining the rich context within videos.
Our extensive experiments over three large benchmarks manifest that COSNet outpe
rforms the current alternatives by a large margin. We will publicly release our
implementation and models.
****

Patch-Based Discriminative Feature Learning for Unsupervised Person Re-Identific
ation
Qize Yang, Hong-Xing Yu, Ancong Wu, Wei-Shi Zheng; Proceedings of the IEEE/CV
F Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3633-3
642
While discriminative local features have been shown effective in solving the per
son re-identification problem, they are limited to be trained on fully pairwise
labelled data which is expensive to obtain. In this work, we overcome this probl
em by proposing a patch-based unsupervised learning framework in order to learn
discriminative feature from patches instead of the whole images. The patch-based
learning leverages similarity between patches to learn a discriminative model.
Specifically, we develop a PatchNet to select patches from the feature map and l
earn discriminative features for these patches. To provide effective guidance fo
r the PatchNet to learn discriminative patch feature on unlabeled datasets, we p
ropose an unsupervised patch-based discriminative feature learning loss. In addi
tion, we design an image-level feature learning loss to leverage all the patch f
eatures of the same image to serve as an image-level guidance for the PatchNet.
Extensive experiments validate the superiority of our method for unsupervised pe
rson re-id. Our code is available at https://github.com/QizeYang/PAUL.
****

SPM-Tracker: Series-Parallel Matching for Real-Time Visual Object Tracking
Guangting Wang, Chong Luo, Zhiwei Xiong, Wenjun Zeng; Proceedings of the IEEE
/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 364
3-3652
The greatest challenge facing visual object tracking is the simultaneous require
ments on robustness and discrimination power. In this paper, we propose a SiamFC
-based tracker, named SPM-Tracker, to tackle this challenge. The basic idea is t
o address the two requirements in two separate matching stages. Robustness is st
rengthened in the coarse matching (CM) stage through generalized training while
discrimination power is enhanced in the fine matching (FM) stage through a dista
nce learning network. The two stages are connected in series as the input propos
als of the FM stage are generated by the CM stage. They are also connected in pa
rallel as the matching scores and box location refinements are fused to generate
the final results. This innovative series-parallel structure takes advantage of
both stages and results in superior performance. The proposed SPM-Tracker, runn
ing at 120fps on GPU, achieves an AUC of 0.687 on OTB-100 and an EAO of 0.434 on
VOT-16, exceeding other real-time trackers by a notable margin.
****

Spatial Fusion GAN for Image Synthesis

Fangneng Zhan, Hongyuan Zhu, Shijian Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3653-3662
Recent advances in generative adversarial networks (GANs) have shown great potentials in realistic image synthesis whereas most existing works address synthesis realism in either appearance space or geometry space but few in both. This paper presents an innovative Spatial Fusion GAN (SF-GAN) that combines a geometry synthesizer and an appearance synthesizer to achieve synthesis realism in both geometry and appearance spaces. The geometry synthesizer learns contextual geometries of background images and transforms and places foreground objects into the background images unanimously. The appearance synthesizer adjust the color, brightness and styles of the foreground objects and embeds them into background images harmoniously, where a guided filter is incorporated for detail preserving. The two synthesizers are inter-connected as mutual references which can be trained end-to-end with little supervision. The SF-GAN has been evaluated in two tasks: (1) realistic scene text image synthesis for training better recognition models; (2) glass and hat wearing for realistic matching glasses and hats with real portraits. Qualitative and quantitative comparisons with the state-of-the-art demonstrate the superiority of the proposed SF-GAN.
****

Text Guided Person Image Synthesis
Xingran Zhou, Siyu Huang, Bin Li, Yingming Li, Jiachen Li, Zhongfei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3663-3672
This paper presents a novel method to manipulate the visual appearance (pose and attribute) of a person image according to natural language descriptions. Our method can be boiled down to two stages: 1) text guided pose generation and 2) visual appearance transferred image synthesis. In the first stage, our method infers a reasonable target human pose based on the text. In the second stage, our method synthesizes a realistic and appearance transferred person image according to the text in conjunction with the target pose. Our method extracts sufficient information from the text and establishes a mapping between the image space and the language space, making generating and editing images corresponding to the description possible. We conduct extensive experiments to reveal the effectiveness of our method, as well as using the VQA Perceptual Score as a metric for evaluating the method. It shows for the first time that we can automatically edit the person image from the natural language descriptions.
****

STGAN: A Unified Selective Transfer Network for Arbitrary Image Attribute Editing
Ming Liu, Yukang Ding, Min Xia, Xiao Liu, Errui Ding, Wangmeng Zuo, Shilei Wen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3673-3682
Arbitrary attribute editing generally can be tackled by incorporating encoder-decoder and generative adversarial networks. However, the bottleneck layer in encoder-decoder usually gives rise to blurry and low quality editing result. And adding skip connections improves image quality at the cost of weakened attribute manipulation ability. Moreover, existing methods exploit target attribute vector to guide the flexible translation to desired target domain. In this work, we suggest to address these issues from selective transfer perspective. Considering that specific editing task is certainly only related to the changed attributes instead of all target attributes, our model selectively takes the difference between target and source attribute vectors as input. Furthermore, selective transfer units are incorporated with encoder-decoder to adaptively select and modify encoder feature for enhanced attribute editing. Experiments show that our method (i.e., STGAN) simultaneously improves attribute manipulation accuracy as well as perception quality, and performs favorably against state-of-the-arts in arbitrary face attribute editing and season translation.
****

Towards Instance-Level Image-To-Image Translation
Zhiqiang Shen, Mingyang Huang, Jianping Shi, Xiangyang Xue, Thomas S. Huang;

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti on (CVPR), 2019, pp. 3683-3692
Unpaired Image-to-image Translation is a new rising and challenging vision probl em that aims to learn a mapping between unaligned image pairs in diverse domains . Recent advances in this field like MUNIT and DRIT mainly focus on disentanglin g content and style/attribute from a given image first, then directly adopting t he global style to guide the model to synthesize new domain images. However, thi s kind of approaches severely incurs contradiction if the target domain images a re content-rich with multiple discrepant objects. In this paper, we present a si mple yet effective instance-aware image-to-image translation approach (INIT), wh ich employs the fine-grained local (instance) and global styles to the target im age spatially. The proposed INIT exhibits three import advantages: (1) the insta nce-level objective loss can help learn a more accurate reconstruction and incor porate diverse attributes of objects; (2) the styles used for target domain of l ocal/global areas are from corresponding spatial regions in source domain, whic h intuitively is a more reasonable mapping; (3) the joint training process can b enefit both fine and coarse granularity and incorporates instance information to improve the quality of global translation. We also collect a large-scale benchm ark for the new instance-level translation task. We observe that our synthetic i mages can even benefit real-world vision tasks like generic object detection.
****

Dense Intrinsic Appearance Flow for Human Pose Transfer
Yining Li, Chen Huang, Chen Change Loy; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3693-3702
We present a novel approach for the task of human pose transfer, which aims at s ynthesizing a new image of a person from an input image of that person and a tar get pose. We address the issues of limited correspondences identified between ke ypoints only and invisible pixels due to self-occlusion. Unlike existing methods , we propose to estimate dense and intrinsic 3D appearance flow to better guide the transfer of pixels between poses. In particular, we wish to generate the 3D flow from just the reference and target poses. Training a network for this purpo se is non-trivial, especially when the annotations for 3D appearance flow are sc arce by nature. We address this problem through a flow synthesis stage. This is achieved by fitting a 3D model to the given pose pair and project them back to t he 2D plane to compute the dense appearance flow for training. The synthesized g round-truths are then used to train a feedforward network for efficient mapping from the input and target skeleton poses to the 3D appearance flow. With the app earance flow, we perform feature warping on the input image and generate a photo realistic image of the target pose. Extensive results on DeepFashion and Market-1501 datasets demonstrate the effectiveness of our approach over existing method s. Our code is available at http://mmlab.ie.cuhk.edu.hk/projects/pose-transfer
****

Depth-Aware Video Frame Interpolation
Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, Ming-Hsuan Y ang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog nition (CVPR), 2019, pp. 3703-3712
Video frame interpolation aims to synthesize nonexistent frames in-between the o riginal frames. While significant advances have been made from the recent deep c onvolutional neural networks, the quality of interpolation is often reduced due to large object motion or occlusion. In this work, we propose a video frame inte rpolation method which explicitly detects the occlusion by exploring the depth i nformation. Specifically, we develop a depth-aware flow projection layer to synt hesize intermediate flows that preferably sample closer objects than farther one s. In addition, we learn hierarchical features to gather contextual information from neighboring pixels. The proposed model then warps the input frames, depth m aps, and contextual features based on the optical flow and local interpolation k ernels for synthesizing the output frame. Our model is compact, efficient, and f ully differentiable. Quantitative and qualitative results demonstrate that the p roposed model performs favorably against state-of-the-art frame interpolation me thods on a wide variety of datasets. The source code and pre-trained model are a

vailable at https://github.com/baowenbo/DAIN.
****
Sliced Wasserstein Generative Models
Jiqing Wu, Zhiwu Huang, Dinesh Acharya, Wen Li, Janine Thoma, Danda Pani Pa udel, Luc Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision a nd Pattern Recognition (CVPR), 2019, pp. 3713-3722
In generative modeling, the Wasserstein distance (WD) has emerged as a useful me tric to measure the discrepancy between generated and real data distributions. U nfortunately, it is challenging to approximate the WD of high-dimensional distri butions. In contrast, the sliced Wasserstein distance (SWD) factorizes high-dime nsional distributions into their multiple one-dimensional marginal distributions and is thus easier to approximate. In this paper, we introduce novel approxima tions of the primal and dual SWD. Instead of using a large number of random proj ections, as it is done by conventional SWD approximation methods, we propose to approximate SWDs with a small number of parameterized orthogonal projections in an end-to-end deep learning fashion. As concrete applications of our SWD approxi mations, we design two types of differentiable SWD blocks to equip modern genera tive frameworks---Auto-Encoders (AE) and Generative Adversarial Networks (GAN). In the experiments, we not only show the superiority of the proposed generative models on standard image synthesis benchmarks, but also demonstrate the state-o f-the-art performance on challenging high resolution image and video generation in an unsupervised manner.
****
Deep Flow-Guided Video Inpainting
Rui Xu, Xiaoxiao Li, Bolei Zhou, Chen Change Loy; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3723-37 32
Video inpainting, which aims at filling in missing regions in a video, remains c hallenging due to the difficulty of preserving the precise spatial and temporal coherence of video contents. In this work we propose a novel flow-guided video i npainting approach. Rather than filling in the RGB pixels of each frame directly , we consider the video inpainting as a pixel propagation problem. We first synt hesize a spatially and temporally coherent optical flow field across video frame s using a newly designed Deep Flow Completion network, then use the synthesized flow fields to guide the propagation of pixels to fill up the missing regions in the video. Specifically, the Deep Flow Competion network follows a coarse-to-fi ne refinement strategy to complete the flow fields, while their quality is furth er improved by hard flow example mining. Following the guide of the completed fl ow fields, the missing video regions can be filled up precisely. Our method is e valuated on DAVIS and YouTubeVOS datasets qualitatively and quantitatively, achi eving the state-of-the-art performance in terms of inpainting quality and speed.
****
Video Generation From Single Semantic Label Map
Junting Pan, Chengyu Wang, Xu Jia, Jing Shao, Lu Sheng, Junjie Yan, Xiaoga ng Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern R ecognition (CVPR), 2019, pp. 3733-3742
This paper proposes the novel task of video generation conditioned on a SINGLE s emantic label map, which provides a good balance between flexibility and quality in the generation process. Different from typical end-to-end approaches, which model both scene content and dynamics in a single step, we propose to decompose this difficult task into two sub-problems. As current image generation methods d o better than video generation in terms of detail, we synthesize high quality co ntent by only generating the first frame. Then we animate the scene based on its semantic meaning to obtain temporally coherent video, giving us excellent resul ts overall. We employ a cVAE for predicting optical flow as a beneficial interme diate step to generate a video sequence conditioned on the initial single frame. A semantic label map is integrated into the flow prediction module to achieve m ajor improvements in the image-to-video generation process. Extensive experiment s on the Cityscapes dataset show that our method outperforms all competing metho ds.

****
Polarimetric Camera Calibration Using an LCD Monitor
Zhixiang Wang, Yinqiang Zheng, Yung-Yu Chuang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3743-3752
It is crucial for polarimetric imaging to accurately calibrate the polarizer angles and the camera response function (CRF) of a polarizing camera. When this polarizing camera is used in a setting of multiview geometric imaging, it is often required to calibrate its intrinsic and extrinsic parameters as well, for which Zhang's calibration method is the most widely used with either a physical checker board, or more conveniently a virtual checker pattern displayed on a monitor. In this paper, we propose to jointly calibrate the polarizer angles and the inverse CRF (ICRF) using a slightly adapted checker pattern displayed on a liquid crystal display (LCD) monitor. Thanks to the lighting principles and the industry standards of the LCD monitors, the polarimetric and radiometric calibration can be significantly simplified, when assisted by the extrinsic parameters estimated from the checker pattern. We present a simple linear method for polarizer angle calibration and a convex method for radiometric calibration, both of which can be jointly refined in a process similar to bundle adjustment. Experiments have verified the feasibility and accuracy of the proposed calibration method.
****
Fully Automatic Video Colorization With Self-Regularization and Diversity
Chenyang Lei, Qifeng Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3753-3761
We present a fully automatic approach to video colorization with self-regularization and diversity. Our model contains a colorization network for video frame colorization and a refinement network for spatiotemporal color refinement. Without any labeled data, both networks can be trained with self-regularized losses defined in bilateral and temporal space. The bilateral loss enforces color consistency between neighboring pixels in a bilateral space and the temporal loss imposes constraints between corresponding pixels in two nearby frames. While video colorization is a multi-modal problem, our method uses a perceptual loss with diversity to differentiate various modes in the solution space. Perceptual experiments demonstrate that our approach outperforms state-of-the-art approaches on fully automatic video colorization.
****
Zoom to Learn, Learn to Zoom
Xuaner Zhang, Qifeng Chen, Ren Ng, Vladlen Koltun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3762-3770
This paper shows that when applying machine learning to digital zoom, it is beneficial to operate on real, RAW sensor data. Existing learning-based super-resolution methods do not use real sensor data, instead operating on processed RGB images. We show that these approaches forfeit detail and accuracy that can be gained by operating on raw data, particularly when zooming in on distant objects. The key barrier to using real sensor data for training is that ground-truth high-resolution imagery is missing. We show how to obtain such ground-truth data via optical zoom and contribute a dataset, SR-RAW, for real-world computational zoom. We use SR-RAW to train a deep network with a novel contextual bilateral loss that is robust to mild misalignment between input and outputs images. The trained network achieves state-of-the-art performance in 4X and 8X computational zoom. We also show that synthesizing sensor data by resampling high-resolution RGB images is an oversimplified approximation of real sensor data and noise, resulting in worse image quality.
****
Single Image Reflection Removal Beyond Linearity
Qiang Wen, Yinjie Tan, Jing Qin, Wenxi Liu, Guoqiang Han, Shengfeng He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3771-3779
Due to the lack of paired data, the training of image reflection removal relies heavily on synthesizing reflection images. However, existing methods model refle

ction as a linear combination model, which cannot fully simulate the real-world scenarios. In this paper, we inject non-linearity into reflection removal from two aspects. First, instead of synthesizing reflection with a fixed combination factor or kernel, we propose to synthesize reflection images by predicting a non-linear alpha blending mask. This enables a free combination of different blurry kernels, leading to a controllable and diverse reflection synthesis. Second, we design a cascaded network for reflection removal with three tasks: predicting the transmission layer, reflection layer, and the non-linear alpha blending mask. The former two tasks are the fundamental outputs, while the latter one being the side output of the network. This side output, on the other hand, making the training a closed loop, so that the separated transmission and reflection layers can be recombined together for training with a reconstruction loss. Extensive quantitative and qualitative experiments demonstrate the proposed synthesis and removal approaches outperforms state-of-the-art methods on two standard benchmarks, as well as in real-world scenarios.

****

Learning to Separate Multiple Illuminants in a Single Image

Zhuo Hui, Ayan Chakrabarti, Kalyan Sunkavalli, Aswin C. Sankaranarayanan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3780-3789

We present a method to separate a single image captured under two illuminants, with different spectra, into the two images corresponding to the appearance of the scene under each individual illuminant. We do this by training a deep neural network to predict the per-pixel reflectance chromaticity of the scene, which we use in conjunction with a previous flash/no-flash image-based separation algorithm to produce the final two output images. We design our reflectance chromaticity network and loss functions by incorporating intuitions from the physics of image formation. We show that this leads to significantly better performance than other single image techniques and even approaches the quality of the two image separation method.

****

Shape Unicode: A Unified Shape Representation

Sanjeev Muralikrishnan, Vladimir G. Kim, Matthew Fisher, Siddhartha Chaudhuri; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3790-3799

3D shapes come in varied representations from a set of points to a set of images, each capturing different aspects of the shape. We propose a unified code for 3D shapes, dubbed Shape Unicode, that imbibes shape cues across these representations into a single code, and a novel framework to learn such a code space for any 3D shape dataset. We discuss this framework as a single go-to training model for any input representation, and demonstrate the effectiveness of the learned code space by applying it directly to common shape analysis tasks -- discriminative and generative. In this work, we use three common representations -- voxel grids, point clouds and multi-view projections -- and combine them into a single code. Note that while we use all three representations at training time, the code can be derived from any single representation during testing. We evaluate this code space on shape retrieval, segmentation and correspondence, and show that the unified code performs better than the individual representations themselves. Additionally, this code space compares quite well to the representation-specific state-of-the-art in these tasks. We also qualitatively discuss linear interpolation between points in this space, by synthesizing from intermediate points.

****

Robust Video Stabilization by Optimization in CNN Weight Space

Jiyang Yu, Ravi Ramamoorthi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3800-3808

We propose a novel robust video stabilization method. Unlike traditional video stabilization techniques that involve complex motion models, we directly model the appearance change of the frames as the dense optical flow field of consecutive frames. We introduce a new formulation of the video stabilization task based on first principles, which leads to a large scale non-convex problem. This problem

is hard to solve, so previous optical flow based approaches have resorted to heuristics. In this paper, we propose a novel optimization routine that transfers this problem into the convolutional neural network parameter domain. While we exploit the general benefits of CNNs, including standard gradient-based optimization techniques, our method is a new approach to using CNNs purely as an optimizer rather than learning from data.Our method trains the CNN from scratch on each specific input example, and intentionally overfits the CNN parameters to produce the best result on the input example. By solving the problem in the CNN weight space rather than directly for image pixels, we make it a viable formulation for video stabilization. Our method produces both visually and quantitatively better results than previous work, and is robust in situations acknowledged as limitations in current state-of-the-art methods.

****

Learning Linear Transformations for Fast Image and Video Style Transfer

Xueting Li, Sifei Liu, Jan Kautz, Ming-Hsuan Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3809-3817

Given a random pair of images, a universal style transfer method extracts the feel from a reference image to synthesize an output based on the look of a content image. Recent algorithms based on second-order statistics, however, are either computationally expensive or prone to generate artifacts due to the trade-off between image quality and runtime performance. In this work, we present an approach for universal style transfer that learns the transformation matrix in a data-driven fashion. Our algorithm is efficient yet flexible to transfer different levels of styles with the same auto-encoder network. It also produces stable video style transfer results due to the preservation of the content affinity. In addition, we propose a linear propagation module to enable a feed-forward network for photo-realistic style transfer. We demonstrate the effectiveness of our approach on three tasks: artistic style, photo-realistic and video style transfer, with comparisons to state-of-the-art methods.

****

Local Detection of Stereo Occlusion Boundaries

Jialiang Wang, Todd Zickler; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3818-3827

Stereo occlusion boundaries are one-dimensional structures in the visual field that separate foreground regions of a scene that are visible to both eyes (binocular regions) from background regions of a scene that are visible to only one eye (monocular regions). Stereo occlusion boundaries often coincide with object boundaries, and localizing them is useful for tasks like grasping, manipulation, and navigation. This paper describes the local signatures for stereo occlusion boundaries that exist in a stereo cost volume, and it introduces a local detector for them based on a simple feedforward network with relatively small receptive fields. The local detector produces better boundaries than many other stereo methods, even without incorporating explicit stereo matching, top-down contextual cues, or single-image boundary cues based on texture and intensity.

****

Bi-Directional Cascade Network for Perceptual Edge Detection

Jianzhong He, Shiliang Zhang, Ming Yang, Yanhu Shan, Tiejun Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3828-3837

Exploiting multi-scale representations is critical to improve edge detection for objects at different scales. To extract edges at dramatically different scales, we propose a Bi-Directional Cascade Network (BDCN) structure, where an individual layer is supervised by labeled edges at its specific scale, rather than directly applying the same supervision to all CNN outputs. Furthermore, to enrich multi-scale representations learned by BDCN, we introduce a Scale Enhancement Module (SEM) which utilizes dilated convolution to generate multi-scale features, instead of using deeper CNNs or explicitly fusing multi-scale edge maps. These new approaches encourage the learning of multi-scale representations in different layers and detect edges that are well delineated by their scales. Learning scale d

edicated layers also results in compact network with a fraction of parameters. We evaluate our method on three datasets, i.e., BSDS500, NYUDv2, and Multicue, and achieve ODS Fmeasure of 0.828, 1.3% higher than current state-of-the art on BSDS500.
****

Single Image Deraining: A Comprehensive Benchmark Analysis
Siyuan Li, Iago Breno Araujo, Wenqi Ren, Zhangyang Wang, Eric K. Tokuda, Roberto Hirata Junior, Roberto Cesar-Junior, Jiawan Zhang, Xiaojie Guo, Xiaochun Cao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3838-3847
We present a comprehensive study and evaluation of existing single image deraining algorithms, using a new large-scale benchmark consisting of both synthetic and real-world rainy images.This dataset highlights diverse data sources and image contents, and is divided into three subsets (rain streak, rain drop, rain and mist), each serving different training or evaluation purposes. We further provide a rich variety of criteria for dehazing algorithm evaluation, ranging from full-reference metrics, to no-reference metrics, to subjective evaluation and the novel task-driven evaluation. Experiments on the dataset shed light on the comparisons and limitations of state-of-the-art deraining algorithms, and suggest promising future directions.
****

Dynamic Scene Deblurring With Parameter Selective Sharing and Nested Skip Connections
Hongyun Gao, Xin Tao, Xiaoyong Shen, Jiaya Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3848-3856
Dynamic Scene deblurring is a challenging low-level vision task where spatially variant blur is caused by many factors, e.g., camera shake and object motion. Recent study has made significant progress. Compared with the parameter independence scheme [19] and parameter sharing scheme [33], we develop the general principle for constraining the deblurring network structure by proposing the generic and effective selective sharing scheme. Inside the subnetwork of each scale, we propose a nested skip connection structure for the nonlinear transformation modules to replace stacked convolution layers or residual blocks. Besides, we build a new large dataset of blurred/sharp image pairs towards better restoration quality. Comprehensive experimental results show that our parameter selective sharing scheme, nested skip connection structure, and the new dataset are all significant to set a new state-of-the-art in dynamic scene deblurring.
****

Events-To-Video: Bringing Modern Computer Vision to Event Cameras
Henri Rebecq, Rene Ranftl, Vladlen Koltun, Davide Scaramuzza; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3857-3866
Event cameras are novel sensors that report brightness changes in the form of asynchronous "events" instead of intensity frames. They have significant advantages over conventional cameras: high temporal resolution, high dynamic range, and no motion blur. Since the output of event cameras is fundamentally different from conventional cameras, it is commonly accepted that they require the development of specialized algorithms to accommodate the particular nature of events. In this work, we take a different view and propose to apply existing, mature computer vision techniques to videos reconstructed from event data. We propose a novel, recurrent neural network to reconstruct videos from a stream of events and train it on a large amount of simulated event data. Our experiments show that our approach surpasses state-of-the-art reconstruction methods by a large margin (> 20%) in terms of image quality. We further apply off-the-shelf computer vision algorithms to videos reconstructed from event data on tasks such as object classification and visual-inertial odometry, and show that this strategy consistently outperforms algorithms that were specifically designed for event data. We believe that our approach opens the door to bringing the outstanding properties of event cameras to an entirely new range of tasks.
****

Feedback Network for Image Super-Resolution
Zhen Li, Jinglei Yang, Zheng Liu, Xiaomin Yang, Gwanggil Jeon, Wei Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3867-3876
Recent advances in image super-resolution (SR) explored the power of deep learning to achieve a better reconstruction performance. However, the feedback mechanism, which commonly exists in human visual system, has not been fully exploited in existing deep learning based image SR methods. In this paper, we propose an image super-resolution feedback network (SRFBN) to refine low-level representations with high-level information. Specifically, we use hidden states in a recurrent neural network (RNN) with constraints to achieve such feedback manner. A feedback block is designed to handle the feedback connections and to generate powerful high-level representations. The proposed SRFBN comes with a strong early reconstruction ability and can create the final high-resolution image step by step. In addition, we introduce a curriculum learning strategy to make the network well suitable for more complicated tasks, where the low-resolution images are corrupted by multiple types of degradation. Extensive experimental results demonstrate the superiority of the proposed SRFBN in comparison with the state-of-the-art methods. Code is avaliable at https://github.com/Paper99/SRFBN_CVPR19.
****
Semi-Supervised Transfer Learning for Image Rain Removal
Wei Wei, Deyu Meng, Qian Zhao, Zongben Xu, Ying Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3877-3886
Single image rain removal is a typical inverse problem in computer vision. The deep learning technique has been verified to be effective for this task and achieved state-of-the-art performance. However, previous deep learning methods need to pre-collect a large set of image pairs with/without synthesized rain for training, which tends to make the neural network be biased toward learning the specific patterns of the synthesized rain, while be less able to generalize to real test samples whose rain types differ from those in the training data. To this issue, this paper firstly proposes a semi-supervised learning paradigm toward this task. Different from traditional deep learning methods which only use supervised image pairs with/without synthesized rains, we further put real rainy images, without need of their clean ones, into the network training process. This is realized by elaborately formulating the residual between an input rainy image and its expected network output (clear image without rain) as a concise mixture of Gaussians distribution. The network is therefore trained to transfer to adapting the real rain pattern domain instead of only the synthesis rain domain, and thus both the short-of-training-sample and bias-to-supervised-sample issues can be evidently alleviated. Experiments on synthetic and real data verify the superiority of our model compared to the state-of-the-arts.
****
EventNet: Asynchronous Recursive Event Processing
Yusuke Sekikawa, Kosuke Hara, Hideo Saito; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3887-3896
Event cameras are bio-inspired vision sensors that mimic retinas to asynchronously report per-pixel intensity changes rather than outputting an actual intensity image at regular intervals. This new paradigm of image sensor offers significant potential advantages; namely, sparse and non-redundant data representation. Unfortunately, however, most of the existing artificial neural network architectures, such as a CNN, require dense synchronous input data, and therefore, cannot make use of the sparseness of the data. We propose EventNet, a neural network designed for real-time processing of asynchronous event streams in a recursive and event-wise manner. EventNet models dependence of the output on tens of thousands of causal events recursively using a novel temporal coding scheme. As a result, at inference time, our network operates in an event-wise manner that is realized with very few sum-of-the-product operations---look-up table and temporal feature aggregation---which enables processing of 1 mega or more events per second on standard CPU. In experiments using real data, we demonstrated the real-time

performance and robustness of our framework.
****

Recurrent Back-Projection Network for Video Super-Resolution
Muhammad Haris, Gregory Shakhnarovich, Norimichi Ukita; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3897-3906
We proposed a novel architecture for the problem of video super-resolution. We integrate spatial and temporal contexts from continuous video frames using a recurrent encoder-decoder module, that fuses multi-frame information with the more traditional, single frame super-resolution path for the target frame. In contrast to most prior work where frames are pooled together by stacking or warping, our model, the Recurrent Back-Projection Network (RBPN) treats each context frame as a separate source of information. These sources are combined in an iterative refinement framework inspired by the idea of back-projection in multiple-image super-resolution. This is aided by explicitly representing estimated inter-frame motion with respect to the target, rather than explicitly aligning frames. We propose a new video super-resolution benchmark, allowing evaluation at a larger scale and considering videos in different motion regimes. Experimental results demonstrate that our RBPN is superior to existing methods on several datasets.
****

Cascaded Partial Decoder for Fast and Accurate Salient Object Detection
Zhe Wu, Li Su, Qingming Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3907-3916
Existing state-of-the-art salient object detection networks rely on aggregating multi-level features of pre-trained convolutional neural networks (CNNs). However, compared to high-level features, low-level features contribute less to performance. Meanwhile, they raise more computational cost because of their larger spatial resolutions. In this paper, we propose a novel Cascaded Partial Decoder (CPD) framework for fast and accurate salient object detection. On the one hand, the framework constructs partial decoder which discards larger resolution features of shallow layers for acceleration. On the other hand, we observe that integrating features of deep layers will obtain relatively precise saliency map. Therefore we directly utilize generated saliency map to recurrently optimize features of deep layers. This strategy efficiently suppresses distractors in the features and significantly improves their representation ability. Experiments conducted on five benchmark datasets exhibit that the proposed model not only achieves state-of-the-art but also runs much faster than existing models. Besides, we apply the proposed framework to optimize existing multi-level feature aggregation models and significantly improve their efficiency and accuracy.
****

A Simple Pooling-Based Design for Real-Time Salient Object Detection
Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, Jianmin Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3917-3926
We solve the problem of salient object detection by investigating how to expand the role of pooling in convolutional neural networks. Based on the U-shape architecture, we first build a global guidance module (GGM) upon the bottom-up pathway, aiming at providing layers at different feature levels the location information of potential salient objects. We further design a feature aggregation module (FAM) to make the coarse-level semantic information well fused with the fine-level features from the top-down path- way. By adding FAMs after the fusion operations in the top-down pathway, coarse-level features from the GGM can be seamlessly merged with features at various scales. These two pooling-based modules allow the high-level semantic features to be progressively refined, yielding detail enriched saliency maps. Experiment results show that our proposed approach can more accurately locate the salient objects with sharpened details and hence substantially improve the performance compared to the previous state-of-the-arts. Our approach is fast as well and can run at a speed of more than 30 FPS when processing a 300x400 image. Code can be found at http://mmcheng.net/poolnet/.
****

Contrast Prior and Fluid Pyramid Integration for RGBD Salient Object Detection
Jia-Xing Zhao, Yang Cao, Deng-Ping Fan, Ming-Ming Cheng, Xuan-Yi Li, Le Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3927-3936
The large availability of depth sensors provides valuable complementary information for salient object detection (SOD) in RGBD images. However, due to the inherent difference between RGB and depth information, extracting features from the depth channel using ImageNet pre-trained backbone models and fusing them with RGB features directly are sub-optimal. In this paper, we utilize contrast prior, which used to be a dominant cue in none deep learning based SOD approaches, into CNNs-based architecture to enhance the depth information. The enhanced depth cues are further integrated with RGB features for SOD, using a novel fluid pyramid integration, which can make better use of multi-scale cross-modal features. Comprehensive experiments on 5 challenging benchmark datasets demonstrate the superiority of the architecture CPFP over 9 state-of-the-art alternative methods.
****
Progressive Image Deraining Networks: A Better and Simpler Baseline
Dongwei Ren, Wangmeng Zuo, Qinghua Hu, Pengfei Zhu, Deyu Meng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3937-3946
Along with the deraining performance improvement of deep networks, their structures and learning become more and more complicated and diverse, making it difficult to analyze the contribution of various network modules when developing new deraining networks. To handle this issue, this paper provides a better and simpler baseline deraining network by considering network architecture, input and output, and loss functions. Specifically, by repeatedly unfolding a shallow ResNet, progressive ResNet (PRN) is proposed to take advantage of recursive computation. A recurrent layer is further introduced to exploit the dependencies of deep features across stages, forming our progressive recurrent network (PReNet). Furthermore, intra-stage recursive computation of ResNet can be adopted in PRN and PReNet to notably reduce network parameters with unsubstantial degradation in deraining performance. For network input and output, we take both stage-wise result and original rainy image as input to each ResNet and finally output the prediction of residual image. As for loss functions, single MSE or negative SSIM losses are sufficient to train PRN and PReNet. Experiments show that PRN and PReNet perform favorably on both synthetic and real rainy images. Considering its simplicity, efficiency and effectiveness, our models are expected to serve as a suitable baseline in future deraining research. The source codes are available at https://github.com/csdwren/PReNet.
****
GSPN: Generative Shape Proposal Network for 3D Instance Segmentation in Point Cloud
Li Yi, Wang Zhao, He Wang, Minhyuk Sung, Leonidas J. Guibas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3947-3956
We introduce a novel 3D object proposal approach named Generative Shape Proposal Network (GSPN) for instance segmentation in point cloud data. Instead of treating object proposal as a direct bounding box regression problem, we take an analysis-by-synthesis strategy and generate proposals by reconstructing shapes from noisy observations in a scene. We incorporate GSPN into a novel 3D instance segmentation framework named Region-based PointNet (R-PointNet) which allows flexible proposal refinement and instance segmentation generation. We achieve state-of-the-art performance on several 3D instance segmentation tasks. The success of GSPN largely comes from its emphasis on geometric understandings during object proposal, greatly reducing proposals with low objectness.
****
Attentive Relational Networks for Mapping Images to Scene Graphs
Mengshi Qi, Weijian Li, Zhengyuan Yang, Yunhong Wang, Jiebo Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3957-3966

Scene graph generation refers to the task of automatically mapping an image into a semantic structural graph, which requires correctly labeling each extracted object and their interaction relationships. Despite the recent success in object detection using deep learning techniques, inferring complex contextual relationships and structured graph representations from visual data remains a challenging topic. In this study, we propose a novel Attentive Relational Network that consists of two key modules with an object detection backbone to approach this problem. The first module is a semantic transformation module utilized to capture semantic embedded relation features, by translating visual features and linguistic features into a common semantic space. The other module is a graph self-attention module introduced to embed a joint graph representation through assigning various importance weights to neighboring nodes. Finally, accurate scene graphs are produced by the relation inference module to recognize all entities and corresponding relations. We evaluate our proposed method on the widely-adopted Visual Genome Dataset, and the results demonstrate the effectiveness and superiority of our model.

****

Relational Knowledge Distillation

Wonpyo Park, Dongju Kim, Yan Lu, Minsu Cho; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3967-3976

Knowledge distillation aims at transferring knowledge acquired in one model (a teacher) to another model (a student) that is typically smaller. Previous approaches can be expressed as a form of training the student to mimic output activations of individual data examples represented by the teacher. We introduce a novel approach, dubbed relational knowledge distillation (RKD), that transfers mutual relations of data examples instead. For concrete realizations of RKD, we propose distance-wise and angle-wise distillation losses that penalize structural differences in relations. Experiments conducted on different tasks show that the proposed method improves educated student models with a significant margin. In particular for metric learning, it allows students to outperform their teachers' performance, achieving the state of the arts on standard benchmark datasets.

****

Compressing Convolutional Neural Networks via Factorized Convolutional Filters

Tuanhui Li, Baoyuan Wu, Yujiu Yang, Yanbo Fan, Yong Zhang, Wei Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3977-3986

This work studies the model compression for deep convolutional neural networks (CNNs) via filter pruning. The workflow of a traditional pruning consists of three sequential stages: pre-training the original model, selecting the pre-trained filters via ranking according to a manually designed criterion (e.g., the norm of filters), and learning the remained filters via fine-tuning. Most existing works follow this pipeline and focus on designing different ranking criteria for filter selection. However, it is difficult to control the performance due to the separation of filter selection and filter learning. In this work, we propose to conduct filter selection and filter learning simultaneously, in a unified model. To this end, we define a factorized convolutional filter (FCF), consisting of a standard real-valued convolutional filter and a binary scalar, as well as a dot-product operator between them. We train a CNN model with factorized convolutional filters (CNN-FCF) by updating the standard filter using back-propagation, while updating the binary scalar using the alternating direction method of multipliers (ADMM) based optimization method. With this trained CNN-FCF model, we only keep the standard filters corresponding to the 1-valued scalars, while all other filters and all binary scalars are discarded, to obtain a compact CNN model. Extensive experiments on CIFAR-10 and ImageNet demonstrate the superiority of the proposed method over state-of-the-art filter pruning methods.

****

On the Intrinsic Dimensionality of Image Representations

Sixue Gong, Vishnu Naresh Boddeti, Anil K. Jain; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3987-3996

This paper addresses the following questions pertaining to the intrinsic dimensi

onality of any given image representation: (i) estimate its intrinsic dimensiona lity, (ii) develop a deep neural network based non-linear mapping, dubbed DeepMD S, that transforms the ambient representation to the minimal intrinsic space, an d (iii) validate the veracity of the mapping through image matching in the intri nsic space. Experiments on benchmark image datasets (LFW, IJB-C and ImageNet-100) reveal that the intrinsic dimensionality of deep neural network representation s is significantly lower than the dimensionality of the ambient features. For in stance, SphereFace's 512-dim face representation and ResNet's 512-dim image repr esentation have an intrinsic dimensionality of 16 and 19 respectively. Further, the DeepMDS mapping is able to obtain a representation of significantly lower di mensionality while maintaining discriminative ability to a large extent, 59.75% TAR @ 0.1% FAR in 16-dim vs 71.26% TAR in 512-dim on IJB-C and a Top-1 accuracy of 77.0% at 19-dim vs 83.4% at 512-dim on ImageNet-100.
****

Part-Regularized Near-Duplicate Vehicle Re-Identification
Bing He, Jia Li, Yifan Zhao, Yonghong Tian; Proceedings of the IEEE/CVF Confe rence on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3997-4005
Vehicle re-identification (Re-ID) has been attracting more interests in computer vision owing to its great contributions in urban surveillance and intelligent t ransportation. With the development of deep learning approaches, vehicle Re-ID s till faces a near-duplicate challenge, which is to distinguish different instanc es with nearly identical appearances. Previous methods simply rely on the global visual features to handle this problem. In this paper, we proposed a simple but efficient part-regularized discriminative feature preserving method which enhan ces the perceptive ability of subtle discrepancies. We further develop a novel f ramework to integrate part constrains with the global Re-ID modules by introduci ng an detection branch. Our framework is trained end-to-end with combined local and global constrains. Specially, without the part-regularized local constrains in inference step, our Re-ID network outperforms the state-of-the-art method by a large margin on large benchmark datasets VehicleID and VeRi-776.
****

Self-Supervised Spatio-Temporal Representation Learning for Videos by Predicting Motion and Appearance Statistics
Jiangliu Wang, Jianbo Jiao, Linchao Bao, Shengfeng He, Yunhui Liu, Wei Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti on (CVPR), 2019, pp. 4006-4015
We address the problem of video representation learning without human-annotated labels. While previous efforts address the problem by designing novel self-super vised tasks using video data, the learned features are merely on a frame-by-fram e basis, which are not applicable to many video analytic tasks where spatio-temp oral features are prevailing. In this paper we propose a novel self-supervised a pproach to learn spatio-temporal features for video representation. Inspired by the success of two-stream approaches in video classification, we propose to lear n visual features by regressing both motion and appearance statistics along spat ial and temporal dimensions, given only the input video data. Specifically, we e xtract statistical concepts (fast-motion region and the corresponding dominant d irection, spatio-temporal color diversity, dominant color, etc.) from simple pat terns in both spatial and temporal domains. Unlike prior puzzles that are even h ard for humans to solve, the proposed approach is consistent with human inherent visual habits and therefore easy to answer. We conduct extensive experiments wi th C3D to validate the effectiveness of our proposed approach. The experiments s how that our approach can significantly improve the performance of C3D when appl ied to video classification tasks. Code is available at https://github.com/laura -wang/video_repres_mas.
****

Classification-Reconstruction Learning for Open-Set Recognition
Ryota Yoshihashi, Wen Shao, Rei Kawakami, Shaodi You, Makoto Iida, Takeshi Naemura; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern R ecognition (CVPR), 2019, pp. 4016-4025
Open-set classification is a problem of handling 'unknown' classes that are not

contained in the training dataset, whereas traditional classifiers assume that o
nly known classes appear in the test environment. Existing open-set classifiers
rely on deep networks trained in a supervised manner on known classes in the tra
ining set; this causes specialization of learned representations to known classe
s and makes it hard to distinguish unknowns from knowns. In contrast, we train n
etworks for joint classification and reconstruction of input data. This enhances
the learned representation so as to preserve information useful for separating
unknowns from knowns, as well as to discriminate classes of knowns. Our novel Cl
assification-Reconstruction learning for Open-Set Recognition (CROSR) utilizes l
atent representations for reconstruction and enables robust unknown detection wi
thout harming the known-class classification accuracy. Extensive experiments rev
eal that the proposed method outperforms existing deep open-set classifiers in m
ultiple standard datasets and is robust to diverse outliers.
****

Emotion-Aware Human Attention Prediction

Macario O. Cordel II, Shaojing Fan, Zhiqi Shen, Mohan S. Kankanhalli; Proceed
ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR
), 2019, pp. 4026-4035

Despite the recent success in face recognition and object classification, in the
field of human gaze prediction, computer models are still struggling to accurat
ely mimic human attention. One main reason is that visual attention is a complex
human behavior influenced by multiple factors, ranging from low-level features
(e.g., color, contrast) to high-level human perception (e.g., objects interactio
ns, object sentiment), making it difficult to model computationally. In this wor
k, we investigate the relation between object sentiment and human attention. We
first introduce a new evaluation metric (AttI) for measuring human attention tha
t focuses on human fixation consensus. A series of empirical data analyses with
AttI indicate that emotion-evoking objects receive attention favor, especially w
hen they co-occur with emotionally-neutral objects, and this favor varies with d
ifferent image complexity. Based on the empirical analyses, we design a deep neu
ral network for human attention prediction which allows the attention bias on em
otion-evoking objects to be encoded in its feature space. Experiments on two ben
chmark datasets demonstrate its superior performance, especially on metrics that
evaluate relative importance of salient regions. This research provides the cle
arest picture to date on how object sentiments influence human attention, and it
makes one of the first attempts to model this phenomenon computationally.
****

Residual Regression With Semantic Prior for Crowd Counting

Jia Wan, Wenhan Luo, Baoyuan Wu, Antoni B. Chan, Wei Liu; Proceedings of the
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp
. 4036-4045

Crowd counting is a challenging task due to factors such as large variations in
crowdedness and severe occlusions. Although recent deep learning based counting
algorithms have achieved a great progress, the correlation knowledge among sampl
es and the semantic prior have not yet been fully exploited. In this paper, a re
sidual regression framework is proposed for crowd counting utilizing the correla
tion information among samples. By incorporating such information into our netwo
rk, we discover that more intrinsic characteristics can be learned by the networ
k which thus generalizes better to unseen scenarios. Besides, we show how to eff
ectively leverage the semantic prior to improve the performance of crowd countin
g. We also observe that the adversarial loss can be used to improve the quality
of predicted density maps, thus leading to an improvement in crowd counting. Exp
eriments on public datasets demonstrate the effectiveness and generalization abi
lity of the proposed method.
****

Context-Reinforced Semantic Segmentation

Yizhou Zhou, Xiaoyan Sun, Zheng-Jun Zha, Wenjun Zeng; Proceedings of the IEEE
/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 404
6-4055

Recent efforts have shown the importance of context on deep convolutional neural

network based semantic segmentation. Among others, the predicted segmentation map (p-map) itself which encodes rich high-level semantic cues (e.g. objects and layout) can be regarded as a promising source of context. In this paper, we propose a dedicated module, Context Net, to better explore the context information in p-maps. Without introducing any new supervisions, we formulate the context learning problem as a Markov Decision Process and optimize it using reinforcement learning during which the p-map and Context Net are treated as environment and agent, respectively. Through adequate explorations, the Context Net selects the information which has long-term benefit for segmentation inference. By incorporating the Context Net with a baseline segmentation scheme, we then propose a Context-reinforced Semantic Segmentation network (CiSS-Net), which is fully end-to-end trainable. Experimental results show that the learned context brings 3.9% absolute improvement on mIoU over the baseline segmentation method, and the CiSS-Net achieves the state-of-the-art segmentation performance on ADE20K, PASCAL-Context and Cityscapes.

****

Adversarial Structure Matching for Structured Prediction Tasks

Jyh-Jing Hwang, Tsung-Wei Ke, Jianbo Shi, Stella X. Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4056-4065

Pixel-wise losses, i.e., cross-entropy or L2, have been widely used in structured prediction tasks as a spatial extension of generic image classification or regression. However, its i.i.d. assumption neglects the structural regularity present in natural images. Various attempts have been made to incorporate structural reasoning mostly through structure priors in a cooperative way where co-occurring patterns are encouraged. We, on the other hand, approach this problem from an opposing angle and propose a new framework, Adversarial Structure Matching (ASM), for training such structured prediction networks via an adversarial process, in which we train a structure analyzer that provides the supervisory signals, the ASM loss. The structure analyzer is trained to maximize ASM loss, or to emphasize recurring multi-scale hard negative structural mistakes usually among co-occurring patterns. On the contrary, the structured prediction network is trained to reduce those mistakes and is thus enabled to distinguish fine-grained structures. As a result, training structured prediction networks using ASM reduces contextual confusion among objects and improves boundary localization. We demonstrate that ASM outperforms its pixel-wise counterpart and commonly used structure priors, GAN, on three different structured prediction tasks, namely, semantic segmentation, monocular depth estimation, and surface normal prediction.

****

Deep Spectral Clustering Using Dual Autoencoder Network

Xu Yang, Cheng Deng, Feng Zheng, Junchi Yan, Wei Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4066-4075

The clustering methods have recently absorbed even-increasing attention in learning and vision. Deep clustering combines embedding and clustering together to obtain optimal embedding subspace for clustering, which can be more effective compared with conventional clustering methods. In this paper, we propose a joint learning framework for discriminative embedding and spectral clustering. We first devise a dual autoencoder network, which enforces the reconstruction constraint for the latent representations and their noisy versions, to embed the inputs into a latent space for clustering. As such the learned latent representations can be more robust to noise. Then the mutual information estimation is utilized to provide more discriminative information from the inputs. Furthermore, a deep spectral clustering method is applied to embed the latent representations into the eigenspace and subsequently clusters them, which can fully exploit the relationship between inputs to achieve optimal clustering results. Experimental results on benchmark datasets show that our method can significantly outperform state-of-the-art clustering approaches.

****

Deep Asymmetric Metric Learning via Rich Relationship Mining

Xinyi Xu, Yanhua Yang, Cheng Deng, Feng Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4076-4085

Learning effective distance metric between data has gained increasing popularity, for its promising performance on various tasks, such as face verification, zero-shot learning, and image retrieval. A major line of researches employs hard data mining, which makes efforts on searching a subset of significant data. However, hard data mining based approaches only rely on a small percentage of data, which is apt to overfitting. This motivates us to propose a novel framework, named deep asymmetric metric learning via rich relationship mining (DAMLRRM), to mine rich relationship under satisfying sampling size. DAMLRRM constructs two asymmetric data streams that are differently structured and of unequal length. The asymmetric structure enables the two data streams to interlace each other, which allows for the informative comparison between new data pairs over iterations. To improve the generalization ability, we further relax the constraint on the intra-class relationship. Rather than greedily connecting all possible positive pairs, DAMLRRM builds a minimum-cost spanning tree within each category to ensure the formation of a connected region. As such there exists at least one direct or indirect path between arbitrary positive pairs to bridge intra-class relevance. Extensive experimental results on three benchmark datasets including CUB-200-2011, Cars196, and Stanford Online Products show that DAMLRRM effectively boosts the performance of existing deep metric learning approaches.

****

Did It Change? Learning to Detect Point-Of-Interest Changes for Proactive Map Updates

Jerome Revaud, Minhyeok Heo, Rafael S. Rezende, Chanmi You, Seong-Gyun Jeong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4086-4095

Maps are an increasingly important tool in our daily lives, yet their rich semantic content still largely depends on manual input. Motivated by the broad availability of geo-tagged street-view images, we propose a new task aiming to make the map update process more proactive. We focus on automatically detecting changes of Points of Interest (POIs), specifically stores or shops of any kind, based on visual input. Faced with the lack of an appropriate benchmark, we build and release a large dataset, captured in two large shopping centers, that comprises 33K geo-localized images and 578 POIs. We then design a generic approach that compares two image sets captured in the same venue at different times and outputs POI changes as a ranked list of map locations. In contrast to logo or franchise recognition approaches, our system does not depend on an external franchise database. It is instead inspired by recent deep metric learning approaches that learn a similarity function fit to the task at hand. We compare various loss functions to learn a metric aligned with the POI change detection goal, and report promising results.

****

Associatively Segmenting Instances and Semantics in Point Clouds

Xinlong Wang, Shu Liu, Xiaoyong Shen, Chunhua Shen, Jiaya Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4096-4105

A 3D point cloud describes the real scene precisely and intuitively. To date how to segment diversified elements in such an informative 3D scene is rarely discussed. In this paper, we first introduce a simple and flexible framework to segment instances and semantics in point clouds simultaneously. Then, we propose two approaches which make the two tasks take advantage of each other, leading to a win-win situation. Specifically, we make instance segmentation benefit from semantic segmentation through learning semantic-aware point-level instance embedding. Meanwhile, semantic features of the points belonging to the same instance are fused together to make more accurate per-point semantic predictions. Our method largely outperforms the state-of-the-art method in 3D instance segmentation along with a significant improvement in 3D semantic segmentation. Code has been made available at: https://github.com/WXinlong/ASIS.

****

Pattern-Affinitive Propagation Across Depth, Surface Normal and Semantic Segment ation
Zhenyu Zhang, Zhen Cui, Chunyan Xu, Yan Yan, Nicu Sebe, Jian Yang; Proceedi ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 2019, pp. 4106-4115
In this paper, we propose a novel Pattern-Affinitive Propagation (PAP) framework to jointly predict depth, surface normal and semantic segmentation. The motivat ion behind it comes from the statistic observation that pattern-affinitive pairs recur much frequently across different tasks as well as within a task. Thus, we can conduct two types of propagations, cross-task propagation and task-specific propagation, to adaptively diffuse those similar patterns. The former integrate s cross-task affinity patterns to adapt to each task therein through the calcula tion on non-local relationships. Next the latter performs an iterative diffusion in the feature space so that the cross-task affinity patterns can be widely-spr ead within the task. Accordingly, the learning of each task can be regularized a nd boosted by the complementary task-level affinities. Extensive experiments dem onstrate the effectiveness and the superiority of our method on the joint three tasks. Meanwhile, we achieve the state-of-the-art or competitive results on the three related datasets, NYUD-v2, SUN-RGBD and KITTI.
****
Scene Categorization From Contours: Medial Axis Based Salience Measures
Morteza Rezanejad, Gabriel Downs, John Wilder, Dirk B. Walther, Allan Jepson , Sven Dickinson, Kaleem Siddiqi; Proceedings of the IEEE/CVF Conference on Co mputer Vision and Pattern Recognition (CVPR), 2019, pp. 4116-4124
The computer vision community has witnessed recent advances in scene categorizat ion from images, with the state of the art systems now achieving impressive reco gnition rates on challenging benchmarks. Such systems have been trained on photo graphs which include color, texture and shading cues. The geometry of shapes and surfaces, as conveyed by scene contours, is not explicitly considered for this task. Remarkably, humans can accurately recognize natural scenes from line drawi ngs, which consist solely of contour-based shape cues. Here we report the first computer vision study on scene categorization of line drawings derived from popu lar databases including an artist scene database, MIT67 and Places365. Specifica lly, we use off-the-shelf pre-trained Convolutional Neural Networks (CNNs) to pe rform scene classification given only contour information as input, and find per formance levels well above chance. We also show that medial-axis based contour s alience methods can be used to select more informative subsets of contour pixels , and that the variation in CNN classification performance on various choices fo r these subsets is qualitatively similar to that observed in human performance. Moreover, when the salience measures are used to weight the contours, we find th at these weights boost our CNN performance above that for unweighted contour inp ut. That is, the medial axis based salience weights appear to add useful informa tion that is not available when CNNs are trained to use contours alone.
****
Unsupervised Image Captioning
Yang Feng, Lin Ma, Wei Liu, Jiebo Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4125-4134
Deep neural networks have achieved great successes on the image captioning task. However, most of the existing models depend heavily on paired image-sentence da tasets, which are very expensive to acquire. In this paper, we make the first at tempt to train an image captioning model in an unsupervised manner. Instead of r elying on manually labeled image-sentence pairs, our proposed model merely requi res an image set, a sentence corpus, and an existing visual concept detector. Th e sentence corpus is used to teach the captioning model how to generate plausibl e sentences. Meanwhile, the knowledge in the visual concept detector is distille d into the captioning model to guide the model to recognize the visual concepts in an image. In order to further encourage the generated captions to be semantic ally consistent with the image, the image and caption are projected into a commo n latent space so that they can reconstruct each other. Given that the existing sentence corpora are mainly designed for linguistic research and are thus with l

ittle reference to image contents, we crawl a large-scale image description corpus of two million natural sentences to facilitate the unsupervised image captioning scenario. Experimental results show that our proposed model is able to produce quite promising results without any caption annotations.
****
Exact Adversarial Attack to Image Captioning via Structured Output Learning With Latent Variables
Yan Xu, Baoyuan Wu, Fumin Shen, Yanbo Fan, Yong Zhang, Heng Tao Shen, Wei Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4135-4144
In this work, we study the robustness of a CNN+RNN based image captioning system being subjected to adversarial noises. We propose to fool an image captioning system to generate some targeted partial captions for an image polluted by adversarial noises, even the targeted captions are totally irrelevant to the image content. A partial caption indicates that the words at some locations in this caption are observed, while words at other locations are not restricted. It is the first work to study exact adversarial attacks of targeted partial captions. Due to the sequential dependencies among words in a caption, we formulate the generation of adversarial noises for targeted partial captions as a structured output learning problem with latent variables. Both the generalized expectation maximization algorithm and structural SVMs with latent variables are then adopted to optimize the problem. The proposed methods generate very successful attacks to three popular CNN+RNN based image captioning models. Furthermore, the proposed attack methods are used to understand the inner mechanism of image captioning systems, providing the guidance to further improve automatic image captioning systems towards human captioning.
****
Cross-Modal Relationship Inference for Grounding Referring Expressions
Sibei Yang, Guanbin Li, Yizhou Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4145-4154
Grounding referring expressions is a fundamental yet challenging task facilitating human-machine communication in the physical world. It locates the target object in an image on the basis of the comprehension of the relationships between referring natural language expressions and the image. A feasible solution for grounding referring expressions not only needs to extract all the necessary information (i.e. objects and the relationships among them) in both the image and referring expressions, but also compute and represent multimodal contexts from the extracted information. Unfortunately, existing work on grounding referring expressions cannot extract multi-order relationships from the referring expressions accurately and the contexts they obtain have discrepancies with the contexts described by referring expressions. In this paper, we propose a Cross-Modal Relationship Extractor (CMRE) to adaptively highlight objects and relationships, that have connections with a given expression, with a cross-modal attention mechanism, and represent the extracted information as a language-guided visual relation graph. In addition, we propose a Gated Graph Convolutional Network (GGCN) to compute multimodal semantic contexts by fusing information from different modes and propagating multimodal information in the structured relation graph. Experiments on various common benchmark datasets show that our Cross-Modal Relationship Inference Network, which consists of CMRE and GGCN, outperforms all existing state-of-the-art methods.
****
What's to Know? Uncertainty as a Guide to Asking Goal-Oriented Questions
Ehsan Abbasnejad, Qi Wu, Qinfeng Shi, Anton van den Hengel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4155-4164
One of the core challenges in Visual Dialogue problems is asking the question that will provide the most useful information towards achieving the required objective. Encouraging an agent to ask the right questions is difficult because we don't know a-priori what information the agent will need to achieve its task, and we don't have an explicit model of what it knows already. We propose a solution

to this problem based on a Bayesian model of the uncertainty in the implicit model maintained by the visual dialogue agent, and in the function used to select an appropriate output. By selecting the question that minimises the predicted regret with respect to this implicit model the agent actively reduces ambiguity. The Bayesian model of uncertainty also enables a principled method for identifying when enough information has been acquired, and an action should be selected. We evaluate our approach on two goal-oriented dialogue datasets, one for visual-based collaboration task and the other for a negotiation-based task. Our uncertainty-aware information-seeking model outperforms its counterparts in these two challenging problems.

****

Iterative Alignment Network for Continuous Sign Language Recognition

Junfu Pu, Wengang Zhou, Houqiang Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4165-4174

In this paper, we propose an alignment network with iterative optimization for weakly supervised continuous sign language recognition. Our framework consists of two modules: a 3D convolutional residual network (3D-ResNet) for feature learning and an encoder-decoder network with connectionist temporal classification (CTC) for sequence modelling. The above two modules are optimized in an alternate way. In the encoder-decoder sequence learning network, two decoders are included, i.e., LSTM decoder and CTC decoder. Both decoders are jointly trained by maximum likelihood criterion with a soft Dynamic Time Warping (soft-DTW) alignment constraint. The warping path, which indicates the possible alignment between input video clips and sign words, is used to fine-tune the 3D-ResNet as training labels with classification loss. After fine-tuning, the improved features are extracted for optimization of encoder-decoder sequence learning network in next iteration. The proposed algorithm is evaluated on two large scale continuous sign language recognition benchmarks, i.e., RWTH-PHOENIX-Weather and CSL. Experimental results demonstrate the effectiveness of our proposed method.

****

Neural Sequential Phrase Grounding (SeqGROUND)

Pelin Dogan, Leonid Sigal, Markus Gross; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4175-4184

We propose an end-to-end approach for phrase grounding in images. Unlike prior methods that typically attempt to ground each phrase independently by building an image-text embedding, our architecture formulates grounding of multiple phrases as a sequential and contextual process. Specifically, we encode region proposals and all phrases into two stacks of LSTM cells, along with so-far grounded phrase-region pairs. These LSTM stacks collectively capture context for grounding of the next phrase. The resulting architecture, which we call SeqGROUND, supports many-to-many matching by allowing an image region to be matched to multiple phrases and vice versa. We show competitive performance on the Flickr30K benchmark dataset and, through ablation studies, validate the efficacy of sequential grounding as well as individual design choices in our model architecture.

****

CLEVR-Ref+: Diagnosing Visual Reasoning With Referring Expressions

Runtao Liu, Chenxi Liu, Yutong Bai, Alan L. Yuille; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4185-4194

Referring object detection and referring image segmentation are important tasks that require joint understanding of visual information and natural language. Yet there has been evidence that current benchmark datasets suffer from bias, and current state-of-the-art models cannot be easily evaluated on their intermediate reasoning process. To address these issues and complement similar efforts in visual question answering, we build CLEVR-Ref+, a synthetic diagnostic dataset for referring expression comprehension. The precise locations and attributes of the objects are readily available, and the referring expressions are automatically associated with functional programs. The synthetic nature allows control over dataset bias (through sampling strategy), and the modular programs enable intermediate reasoning ground truth without human annotators. In addition to evaluating

several state-of-the-art models on CLEVR-Ref+, we also propose IEP-Ref, a module network approach that significantly outperforms other models on our dataset. In particular, we present two interesting and important findings using IEP-Ref: (1) the module trained to transform feature maps into segmentation masks can be attached to any intermediate module to reveal the entire reasoning process step-by-step; (2) even if all training data has at least one object referred, IEP-Ref can correctly predict no-foreground when presented with false-premise referring expressions. To the best of our knowledge, this is the first direct and quantitative proof that neural modules behave in the way they are intended. We will release data and code for CLEVR-Ref+.
****

Describing Like Humans: On Diversity in Image Captioning
Qingzhong Wang, Antoni B. Chan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4195-4203
Recently, the state-of-the-art models for image captioning have overtaken human performance based on the most popular metrics, such as BLEU, METEOR, ROUGE and CIDEr. Does this mean we have solved the task of image captioning The above metrics only measure the similarity of the generated caption to the human annotations, which reflects its accuracy. However, an image contains many concepts and multiple levels of detail, and thus there is a variety of captions that express different concepts and details that might be interesting for different humans. Therefore only evaluating accuracy is not sufficient for measuring the performance of captioning models --- the diversity of the generated captions should also be considered. In this paper, we proposed a new metric for measuring the diversity of image captions, which is derived from latent semantic analysis and kernelized to use CIDEr similarity. We conduct extensive experiments to re-evaluate recent captioning models in the context of both diversity and accuracy. We find that there is still a large gap between the model and human performance in terms of both accuracy and diversity, and the models that have optimized accuracy (CIDEr) have low diversity. We also show that balancing the cross-entropy loss and CIDEr reward in reinforcement learning during training can effectively control the trade off between diversity and accuracy of the generated captions.
****

MSCap: Multi-Style Image Captioning With Unpaired Stylized Text
Longteng Guo, Jing Liu, Peng Yao, Jiangwei Li, Hanqing Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4204-4213
In this paper, we propose an adversarial learning network for the task of multi-style image captioning (MSCap) with a standard factual image caption dataset and a multi-stylized language corpus without paired images. How to learn a single model for multi-stylized image captioning with unpaired data is a challenging and necessary task, whereas rarely studied in previous works. The proposed framework mainly includes four contributive modules following a typical image encoder. First, a style dependent caption generator to output a sentence conditioned on an encoded image and a specified style. Second, a caption discriminator is presented to distinguish the input sentence to be real or not. The discriminator and the generator are trained in an adversarial manner to enable more natural and human-like captions. Third, a style classifier is employed to discriminate the specific style of the input sentence. Besides, a back-translation module is designed to enforce the generated stylized captions are visually grounded, with the intuition of the cycle consistency for factual caption and stylized caption. We enable an end-to-end optimization of the whole model with differentiable softmax approximation. At last, we conduct comprehensive experiments using a combined dataset containing four caption styles to demonstrate the outstanding performance of our proposed method.
****

CRAVES: Controlling Robotic Arm With a Vision-Based Economic System
Yiming Zuo, Weichao Qiu, Lingxi Xie, Fangwei Zhong, Yizhou Wang, Alan L. Yuille; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4214-4223

Training a robotic arm to accomplish real-world tasks has been attracting increasing attention in both academia and industry. This work discusses the role of computer vision algorithms in this field. We focus on low-cost arms on which no sensors are equipped and thus all decisions are made upon visual recognition, e.g., real-time 3D pose estimation. This requires annotating a lot of training data, which is not only time-consuming but also laborious. In this paper, we present an alternative solution, which uses a 3D model to create a large number of synthetic data, trains a vision model in this virtual domain, and applies it to real-world images after domain adaptation. To this end, we design a semi-supervised approach, which fully leverages the geometric constraints among keypoints. We apply an iterative algorithm for optimization. Without any annotations on real images, our algorithm generalizes well and produces satisfying results on 3D pose estimation, which is evaluated on two real-world datasets. We also construct a vision-based control system for task accomplishment, for which we train a reinforcement learning agent in a virtual environment and apply it to the real-world. Moreover, our approach, with merely a 3D model being required, has the potential to generalize to other types of multi-rigid-body dynamic systems.

****

Networks for Joint Affine and Non-Parametric Image Registration

Zhengyang Shen, Xu Han, Zhenlin Xu, Marc Niethammer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4224-4233

We introduce an end-to-end deep-learning framework for 3D medical image registration. In contrast to existing approaches, our framework combines two registration methods: an affine registration and a vector momentum-parameterized stationary velocity field (vSVF) model. Specifically, it consists of three stages. In the first stage, a multi-step affine network predicts affine transform parameters. In the second stage, we use a U-Net-like network to generate a momentum, from which a velocity field can be computed via smoothing. Finally, in the third stage, we employ a self-iterable map-based vSVF component to provide a non-parametric refinement based on the current estimate of the transformation map. Once the model is trained, a registration is completed in one forward pass. To evaluate the performance, we conducted longitudinal and cross-subject experiments on 3D magnetic resonance images (MRI) of the knee of the Osteoarthritis Initiative (OAI) dataset. Results show that our framework achieves comparable performance to state-of-the-art medical image registration approaches, but it is much faster, with a better control of transformation regularity including the ability to produce approximately symmetric transformations, and combining affine as well as non-parametric registration.

****

Learning Shape-Aware Embedding for Scene Text Detection

Zhuotao Tian, Michelle Shu, Pengyuan Lyu, Ruiyu Li, Chao Zhou, Xiaoyong Shen, Jiaya Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4234-4243

We address the problem of detecting scene text in arbitrary shapes, which is a challenging task due to the high variety and complexity of the scene. Specifically, we treat text detection as instance segmentation and propose a segmentation-based framework, which extracts each text instance as an independent connected component. To distinguish different text instances, our method maps pixels onto an embedding space where pixels belonging to the same text are encouraged to appear closer to each other and vise versa. In addition, we introduce a Shape-Aware Loss to make training adaptively accommodate various aspect ratios of text instances and the tiny gaps among them, and a new post-processing pipeline to yield precise bounding box predictions. Experimental results on three challenging datasets (ICDAR15, MSRA-TD500 and CTW1500) demonstrate the effectiveness of our work.

****

Learning to Film From Professional Human Motion Videos

Chong Huang, Chuan-En Lin, Zhenyu Yang, Yan Kong, Peng Chen, Xin Yang, Kwang-Ting Cheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4244-4253

We investigate the problem of 6 degrees of freedom (DOF) camera planning for filming professional human motion videos using a camera drone. Existing methods either plan motions for only a pan-tilt-zoom (PTZ) camera, or adopt ad-hoc solutions without carefully considering the impact of video contents and previous camera motions on the future camera motions. As a result, they can hardly achieve satisfactory results in our drone cinematography task. In this study, we propose a learning-based framework which incorporates the video contents and previous camera motions to predict the future camera motions that enable the capture of professional videos. Specifically, the inputs of our framework are video contents which are represented using subject-related feature based on 2D skeleton and scene-related features extracted from background RGB images, and camera motions which are represented using optical flows. The correlation between the inputs and output future camera motions are learned via a sequence-to-sequence convolutional long short-term memory (Seq2Seq ConvLSTM) network from a large set of video clips. We deploy our approach to a real drone cinematography system by first predicting the future camera motions, and then converting them to the drone's control commands via an odometer. Our experimental results on extensive datasets and showcases exhibit significant improvements in our approach over conventional baselines and our approach can successfully mimic the footage of a professional cameraman.

****

Pay Attention! - Robustifying a Deep Visuomotor Policy Through Task-Focused Visual Attention

Pooya Abolghasemi, Amir Mazaheri, Mubarak Shah, Ladislau Boloni; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4254-4262

Several recent studies have demonstrated the promise of deep visuomotor policies for robot manipulator control. Despite impressive progress, these systems are known to be vulnerable to physical disturbances, such as accidental or adversarial bumps that make them drop the manipulated object. They also tend to be distracted by visual disturbances such as objects moving in the robot's field of view, even if the disturbance does not physically prevent the execution of the task. In this paper, we propose an approach for augmenting a deep visuomotor policy trained through demonstrations with Task Focused visual Attention (TFA). The manipulation task is specified with a natural language text such as "move the red bowl to the left". This allows the visual attention component to concentrate on the current object that the robot needs to manipulate. We show that even in benign environments, the TFA allows the policy to consistently outperform a variant with no attention mechanism. More importantly, the new policy is significantly more robust: it regularly recovers from severe physical disturbances (such as bumps causing it to drop the object) from which the baseline policy, i.e. with no visual attention, almost never recovers. In addition, we show that the proposed policy performs correctly in the presence of a wide class of visual disturbances, exhibiting a behavior reminiscent of human selective visual attention experiments.

****

Deep Blind Video Decaptioning by Temporal Aggregation and Recurrence

Dahun Kim, Sanghyun Woo, Joon-Young Lee, In So Kweon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4263-4272

Blind video decaptioning is a problem of automatically removing text overlays and inpainting the occluded parts in videos without any input masks. While recent deep learning based inpainting methods deal with a single image and mostly assume that the positions of the corrupted pixels are known, we aim at automatic text removal in video sequences without mask information. In this paper, we propose a simple yet effective framework for fast blind video decaptioning. We construct an encoder-decoder model, where the encoder takes multiple source frames that can provide visible pixels revealed from the scene dynamics. These hints are aggregated and fed into the decoder. We apply a residual connection from the input frame to the decoder output to enforce our network to focus on the corrupted regions only. Our proposed model was ranked in the first place in the ECCV Chalearn

2018 LAP Inpainting Competition Track2: Video decaptioning. In addition, we furt
her improve this strong model by applying a recurrent feedback. The recurrent fe
edback not only enforces temporal coherence but also provides strong clues on wh
ere the corrupted pixels are. Both qualitative and quantitative experiments demo
nstrate that our full model produces accurate and temporally consistent video re
sults in real time (50+ fps).
****