3D-RCNN: Instance-Level 3D Object Reconstruction via Render-and-Compare
Abhijit Kundu, Yin Li, James M. Rehg; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3559-3568

We present a fast inverse-graphics framework for instance-level 3D scene understanding. We train a deep convolutional network that learns to map image regions to the full 3D shape and pose of all object instances in the image. Our method produces a compact 3D representation of the scene, which can be readily used for applications like autonomous driving. Many traditional 2D vision outputs, like instance segmentations and depth-maps, can be obtained by simply rendering our output 3D scene model. We exploit class-specific shape priors by learning a low dimensional shape-space from collections of CAD models. We present novel representations of shape and pose, that strive towards better 3D equivariance and generalization. In order to exploit rich supervisory signals in the form of 2D annotations like segmentation, we propose a differentiable Render-and-Compare loss that allows 3D shape and pose to be learned with 2D supervision. We evaluate our method on the challenging real-world datasets of Pascal3D+ and KITTI, where we achieve state-of-the-art results.

****

Fast and Furious: Real Time End-to-End 3D Detection, Tracking and Motion Forecasting With a Single Convolutional Net
Wenjie Luo, Bin Yang, Raquel Urtasun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3569-3577

In this paper we propose a novel deep neural network that is able to jointly reason about 3D detection, tracking and motion forecasting given data captured by a 3D sensor. By jointly reasoning about these tasks, our holistic approach is more robust to occlusion as well as sparse data at range. Our approach performs 3D convolutions across space and time over a bird's eye view representation of the 3D world, which is very efficient in terms of both memory and computation. Our experiments on a new very large scale dataset captured in several north american cities, show that we can outperform the state-of-the-art by a large margin. Importantly, by sharing computation we can perform all tasks in as little as 30 ms.

****

An Analysis of Scale Invariance in Object Detection - SNIP
Bharat Singh, Larry S. Davis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3578-3587

An analysis of different techniques for recognizing and detecting objects under extreme scale variation is presented. Scale specific and scale invariant design of detectors are compared by training them with different configurations of input data. By evaluating the performance of different network architectures for classifying small objects on ImageNet, we show that CNNs are not robust to changes in scale. Based on this analysis, we propose to train and test detectors on the same scales of an image-pyramid. Since small and large objects are difficult to recognize at smaller and larger scales respectively, we present a novel training scheme called Scale Normalization for Image Pyramids (SNIP) which selectively back-propagates the gradients of object instances of different sizes as a function of the image scale. On the COCO dataset, our single model performance is 45.7% and an ensemble of 3 networks obtains an mAP of 48.3%. We use off-the-shelf ImageNet-1000 pre-trained models and only train with bounding box supervision. Our submission won the Best Student Entry in the COCO 2017 challenge. Code will be made available at url{http://bit.ly/2yXVg4c}.

****

Relation Networks for Object Detection
Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, Yichen Wei; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3588-3597

Although it is well believed for years that modeling relations between objects would help object recognition, there has not been evidence that the idea is working in the deep learning era. All state-of-the-art object detection systems still rely on recognizing object instances ■extbf{individually}, without exploiting their relations during learning. This work proposes an object relation module. It processes a set of objects ■extbf{simultaneously} through interaction between their appearance feature and geometry, thus allowing modeling of their relations. It is lightweight and in-place. It does not require additional supervision and is easy to embed in existing networks. It is shown effective on improving object recognition and duplicate removal steps in the modern object detection pipeline. It verifies the efficacy of modeling object relations in CNN based detection. It gives rise to the ■extbf{first fully end-to-end object detector}.

****

Zero-Shot Sketch-Image Hashing

Yuming Shen, Li Liu, Fumin Shen, Ling Shao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3598-3607

Recent studies show that large-scale sketch-based image retrieval (SBIR) can be efficiently tackled by cross-modal binary representation learning methods, where Hamming distance matching significantly speeds up the process of similarity search. Providing training and test data subjected to a fixed set of pre-defined categories, the cutting-edge SBIR and cross-modal hashing works obtain acceptable retrieval performance. However, most of the existing methods fail when the categories of query sketches have never been seen during training. In this paper, the above problem is briefed as a novel but realistic zero-shot SBIR hashing task. We elaborate the challenges of this special task and accordingly propose a zero-shot sketch-image hashing (ZSIH) model. An end-to-end three-network architecture is built, two of which are treated as the binary encoders. The third network mitigates the sketch-image heterogeneity and enhances the semantic relations among data by utilizing the Kronecker fusion layer and graph convolution, respectively. As an important part of ZSIH, we formulate a generative hashing scheme in reconstructing semantic knowledge representations for zero-shot retrieval. To the best of our knowledge, ZSIH is the first zero-shot hashing work suitable for SBIR and cross-modal search. Comprehensive experiments are conducted on two extended datasets, i.e., Sketchy and TU-Berlin with a novel zero-shot train-test split. The proposed model remarkably outperforms related works.

****

VizWiz Grand Challenge: Answering Visual Questions From Blind People

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, Jeffrey P. Bigham; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3608-3617

The study of algorithms to automatically answer visual questions currently is motivated by visual question answering (VQA) datasets constructed in artificial VQA settings. We propose VizWiz, the first goal-oriented VQA dataset arising from a natural VQA setting. VizWiz consists of 31,000 visual questions originating from blind people who each took a picture using a mobile phone and recorded a spoken question about it, together with 10 crowdsourced answers per visual question. VizWiz differs from the many existing VQA datasets because (1) images are captured by blind photographers and so are often poor quality, (2) questions are spoken and so are more conversational, and (3) often visual questions cannot be answered. Evaluation of modern algorithms for answering visual questions and deciding if a visual question is answerable reveals that VizWiz is a challenging dataset. We introduce this dataset to encourage a larger community to develop more generalized algorithms that can assist blind people.

****

Divide and Grow: Capturing Huge Diversity in Crowd Images With Incrementally Growing CNN

Deepak Babu Sam, Neeraj N. Sajjan, R. Venkatesh Babu, Mukundhan Srinivasan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3618-3626

Automated counting of people in crowd images is a challenging task. The major difficulty stems from the large diversity in the way people appear in crowds. In fact, features available for crowd discrimination largely depend on the crowd density to the extent that people are only seen as blobs in a highly dense scene. We tackle this problem with a growing CNN which can progressively increase its ca pacity to account for the wide variability seen in crowd scenes. Our model starts from a base CNN density regressor, which is trained in equivalence on all types of crowd images. In order to adapt with the huge diversity, we create two child regressors which are exact copies of the base CNN. A differential training procedure divides the dataset into two clusters and fine-tunes the child networks on their respective specialties. Consequently, without any hand-crafted criteria for forming specialties, the child regressors become experts on certain types of crowds. The child networks are again split recursively, creating two experts at every division. This hierarchical training leads to a CNN tree, where the child regressors are more fine experts than any of their parents. The leaf nodes are taken as the final experts and a classifier network is then trained to predict the correct specialty for a given test image patch. The proposed model achieves higher count accuracy on major crowd datasets. Further, we analyse the characteristics of specialties mined automatically by our method.

****

Structured Set Matching Networks for One-Shot Part Labeling

Jonghyun Choi, Jayant Krishnamurthy, Aniruddha Kembhavi, Ali Farhadi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3627-3636

Diagrams often depict complex phenomena and serve as a good test bed for visual and textual reasoning. However, understanding diagrams using natural image understanding approaches requires large training datasets of diagrams, which are very hard to obtain. Instead, this can be addressed as a matching problem either between labeled diagrams, images or both. This problem is very challenging since the absence of significant color and texture renders local cues ambiguous and requires global reasoning. We consider the problem of one-shot part labeling: labeling multiple parts of an object in a target image given only a single source image of that category. For this set-to-set matching problem, we introduce the Structured Set Matching Network (SSMN), a structured prediction model that incorporates convolutional neural networks. The SSMN is trained using global normalization to maximize local match scores between corresponding elements and a global consistency score among all matched elements, while also enforcing a matching constraint between the two sets. The SSMN significantly outperforms several strong baselines on three label transfer scenarios: diagram-to-diagram, evaluated on a new diagram dataset of over 200 categories; image-to-image, evaluated on a dataset built on top of the Pascal Part Dataset; and image-to-diagram, evaluated on transferring labels across these datasets.

****

Self-Supervised Learning of Geometrically Stable Features Through Probabilistic Introspection

David Novotny, Samuel Albanie, Diane Larlus, Andrea Vedaldi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3637-3645

Self-supervision can dramatically cut back the amount of manually-labelled data required to train deep neural networks. While self-supervision has usually been considered for tasks such as image classification, in this paper we aim at extending it to geometry-oriented tasks such as semantic matching and part detection. We do so by building on several recent ideas in unsupervised landmark detection. Our approach learns dense distinctive visual descriptors from an unlabeled dataset of images using synthetic image transformations. It does so by means of a robust probabilistic formulation that can introspectively determine which image regions are likely to result in stable image matching. We show empirically that a network pre-trained in this manner requires significantly less supervision to learn semantic object parts compared to numerous pre-training alternatives. We also show that the pre-trained representation is excellent for semantic object matching.

****

Link and Code: Fast Indexing With Graphs and Compact Regression Codes

Matthijs Douze, Alexandre Sablayrolles, Hervé Jégou; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3646-3654

Similarity search approaches based on graph walks have recently attained outstanding speed-accuracy trade-offs, taking aside the memory requirements. In this paper, we revisit these approaches by considering, additionally, the memory constraint required to index billions of images on a single server. This leads us to propose a method based both on graph traversal and compact representations. We encode the indexed vectors using quantization and exploit the graph structure to refine the similarity estimation.   In essence, our method takes the best of these two worlds: the search strategy is based on nested graphs, thereby providing high precision with a relatively small set of comparisons. At the same time it offers a significant memory compression. As a result, our approach outperforms the state of the art on operating points considering 64--128 bytes per vector, as demonstrated by our results on two billion-scale public benchmarks.

****

Textbook Question Answering Under Instructor Guidance With Memory Networks

Juzheng Li, Hang Su, Jun Zhu, Siyu Wang, Bo Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3655-3663

Textbook Question Answering (TQA) is a task to choose the most proper answers by reading a multi-modal context of abundant essays and images. TQA serves as a favorable test bed for visual and textual reasoning. However, most of the current methods are incapable of reasoning over the long contexts and images. To address this issue, we propose a novel approach of Instructor Guidance with Memory Networks (IGMN) which conducts the TQA task by finding contradictions between the candidate answers and their corresponding context. We build the Contradiction Entity-Relationship Graph (CERG) to extend the passage-level multi-modal contradictions to an essay level. The machine thus performs as an instructor to extract the essay-level contradictions as the Guidance. Afterwards, we exploit the memory networks to capture the information in the Guidance, and use the attention mechanisms to jointly reason over the global features of the multi-modal input. Extensive experiments demonstrate that our method outperforms the state-of-the-arts on the TQA dataset. The source code is available at https://github.com/freerailway/igmn.

****

Unsupervised Deep Generative Adversarial Hashing Network

Kamran Ghasedi Dizaji, Feng Zheng, Najmeh Sadoughi, Yanhua Yang, Cheng Deng, Heng Huang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3664-3673

Unsupervised deep hash functions have not shown satisfactory improvements against the shallow alternatives, and usually, require supervised pretraining to avoid getting stuck in bad local minima. In this paper, we propose a deep unsupervised hashing function, called HashGAN, which outperforms unsupervised hashing models with significant margins without any supervised pretraining. HashGAN consists of three networks, a generator, a discriminator and an encoder. By sharing the parameters of the encoder and discriminator, we benefit from the adversarial loss as a data dependent regularization in training our deep hash function. Moreover, a novel loss function is introduced for hashing real images, resulting in minimum entropy, uniform frequency, consistent and independent hash bits. Furthermore, we train the generator conditioning on random binary inputs and also use these binary variables in a triplet ranking loss for improving hash codes. In our experiments, HashGAN outperforms the previous unsupervised hash functions in image retrieval and achieves the state-of-the-art performance in image clustering. We also provide an ablation study, showing the contribution of each component in our loss function.

****

Vision-and-Language Navigation: Interpreting Visually-Grounded Navigation Instructions in Real Environments

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, Anton van den Hengel; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3674-3683

A robot that can carry out a natural-language instruction has been a dream since before the Jetsons cartoon series imagined a life of leisure mediated by a flee t of attentive robot helpers. It is a dream that remains stubbornly distant. However, recent advances in vision and language methods have made incredible progress in closely related areas. This is significant because a robot interpreting a natural-language navigation instruction on the basis of what it sees is carrying out a vision and language process that is similar to Visual Question Answering. Both tasks can be interpreted as visually grounded sequence-to-sequence translation problems, and many of the same methods are applicable. To enable and encourage the application of vision and language methods to the problem of interpreting visually-grounded navigation instructions, we present the Matterport3D Simulator -- a large-scale reinforcement learning environment based on real imagery. Using this simulator, which can in future support a range of embodied vision and language tasks, we provide the first benchmark dataset for visually-grounded natural language navigation in real buildings -- the Room-to-Room (R2R) dataset.

****

DenseASPP for Semantic Segmentation in Street Scenes

Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, Kuiyuan Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3684-3692

Semantic image segmentation is a basic street scene understanding task in autonomous driving, where each pixel in a high resolution image is categorized into a set of semantic labels. Unlike other scenarios, objects in autonomous driving scene exhibit very large scale changes, which poses great challenges for high-level feature representation in a sense that multi-scale information must be correctly encoded. To remedy this problem, atrous convolutioncite{Deeplabv1} was introduced to generate features with larger receptive fields without sacrificing spatial resolution. Built upon atrous convolution, Atrous Spatial Pyramid Pooling (ASPP)cite{Deeplabv2} was proposed to concatenate multiple atrous-convolved features using different dilation rates into a final feature representation. Although ASPP is able to generate multi-scale features, we argue the feature resolution in the scale-axis is not dense enough for the autonomous driving scenario. To this end, we propose Densely connected Atrous Spatial Pyramid Pooling (DenseASPP), which connects a set of atrous convolutional layers in a dense way, such that it generates multi-scale features that not only cover a larger scale range, but also cover that scale range densely, without significantly increasing the model size. We evaluate DenseASPP on the street scene benchmark Cityscapescite{Cityscapes} and achieve state-of-the-art performance.

****

Efficient Optimization for Rank-Based Loss Functions

Pritish Mohapatra, Michal Rolínek, C.V. Jawahar, Vladimir Kolmogorov, M. Pawan Kumar; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3693-3701

The accuracy of information retrieval systems is often measured using complex loss functions such as the average precision (AP) or the normalized discounted cumulative gain (NDCG). Given a set of positive and negative samples, the parameters of a retrieval system can be estimated by minimizing these loss functions. However, the non-differentiability and non-decomposability of these loss functions does not allow for simple gradient based optimization algorithms. This issue is generally circumvented by either optimizing a structured hinge-loss upper bound to the loss function or by using asymptotic methods like the direct-loss minimization framework. Yet, the high computational complexity of loss-augmented inference, which is necessary for both the frameworks, prohibits its use in large training data sets. To alleviate this deficiency, we present a novel quicksort flavored algorithm for a large class of non-decomposable loss functions. We provide a complete characterization of the loss functions that are amenable to our algorithm, and show that it includes both AP and NDCG based loss functions. Furthermore, we prove that no comparison based algorithm can improve upon the computational complexity of our approach asymptotically. We demonstrate the effectiveness of our approach in the context of optimizing the structured hinge loss upper bound of AP and NDCG loss for learning models for a variety of vision tasks. We show that our approach provides significantly better results than simpler decomposabl e loss functions, while requiring a comparable training time.

****

Wasserstein Introspective Neural Networks

Kwonjoon Lee, Weijian Xu, Fan Fan, Zhuowen Tu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3702-3711

We present Wasserstein introspective neural networks (WINN) that are both a generator and a discriminator within a single model. WINN provides a significant improvement over the recent introspective neural networks (INN) method by enhancing INN's generative modeling capability. WINN has three interesting properties: (1) A mathematical connection between the formulation of the INN algorithm and that of Wasserstein generative adversarial networks (WGAN) is made. (2) The explicit adoption of the Wasserstein distance into INN results in a large enhancement to INN, achieving compelling results even with a single classifier --- e.g., providing nearly a 20 times reduction in model size over INN for unsupervised generative modeling. (3) When applied to supervised classification, WINN also gives rise to improved robustness against adversarial examples in terms of the error reduction. In the experiments, we report encouraging results on unsupervised learning problems including texture, face, and object modeling, as well as a supervised classification task against adversarial attacks.

****

Taskonomy: Disentangling Task Transfer Learning

Amir R. Zamir, Alexander Sax, William Shen, Leonidas J. Guibas, Jitendra Malik, Silvio Savarese; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3712-3722

Do visual tasks have a relationship, or are they unrelated? For instance, could having surface normals simplify estimating the depth of an image? Intuition answers these questions positively, implying existence of a structure among visual tasks. Knowing this structure has notable uses; it is the concept underlying transfer learning and, for example, can provide a principled way for reusing supervision among related tasks, finding what tasks transfer well to an arbitrary target task, or solving many tasks in one system without piling up the complexity.

This paper proposes a fully computational approach for finding the structure of the space of visual tasks. This is done via a sampled dictionary of twenty six 2D, 2.5D, 3D, and semantic tasks, and modeling their (1st and higher order) transfer dependencies in a latent space. The product can be viewed as a computational taxonomic map for task transfer learning. We study the consequences of this structure, e.g. the nontrivial emerged relationships, and exploit them to reduce the demand for labeled data. For example, we show that the total number of labeled datapoints needed for solving a set of 10 tasks can be reduced by roughly 2/3 while keeping the performance nearly the same. Users can employ a provided Binary Integer Programming solver that leverages the taxonomy to find efficient supervision policies for their own use cases.

****

Maximum Classifier Discrepancy for Unsupervised Domain Adaptation

Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, Tatsuya Harada; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3723-3732

In this work, we present a method for unsupervised domain adaptation. Many adversarial learning methods train domain classifier networks to distinguish the features as either a source or target and train a feature generator network to mimic the discriminator. Two problems exist with these methods. First, the domain classifier only tries to distinguish the features as a source or target and thus does not consider task-specific decision boundaries between classes. Therefore, a trained generator can generate ambiguous features near class boundaries. Second, these methods aim to completely match the feature distributions between different domains, which is difficult because of each domain's characteristics. To solve these problems, we introduce a new approach that attempts to align distributions of source and target by utilizing the task-specific decision boundaries. We propose to maximize the discrepancy between two classifiers' outputs to detect target samples that are far from the support of the source. A feature generator learns to generate target features near the support to minimize the discrepancy. Our method outperforms other methods on several datasets of image classification and semantic segmentation. The codes are available at url{https://github.com/mil-tokyo/MCD_DA}

****

Unsupervised Feature Learning via Non-Parametric Instance Discrimination
Zhirong Wu, Yuanjun Xiong, Stella X. Yu, Dahua Lin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3733-3742
Neural net classifiers trained on data with annotated class labels can also capture apparent visual similarity among categories without being directed to do so. We study whether this observation can be extended beyond the conventional domain of supervised learning: Can we learn a good feature representation that captures apparent similarity among instances, instead of classes, by merely asking the feature to be discriminative of individual instances? We formulate this intuition as a non-parametric classification problem at the instance-level, and use noise-contrastive estimation to tackle the computational challenges imposed by the large number of instance classes. Our experimental results demonstrate that, under unsu- pervised learning settings, our method surpasses the state-of-the-art on ImageNet classification by a large margin. Our method is also remarkable for consistently improving test performance with more training data and better network architectures. By fine-tuning the learned feature, we further obtain competitive results for semi-supervised learning and object detection tasks. Our non-parametric model is highly compact: With 128 features per image, our method requires only 600MB storage for a million images, enabling fast nearest neighbour retrieval at the run time.

****

Multi-Task Adversarial Network for Disentangled Feature Learning
Yang Liu, Zhaowen Wang, Hailin Jin, Ian Wassell; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3743-3751
We address the problem of image feature learning for the applications where multiple factors exist in the image generation process and only some factors are of our interest. We present a novel multi-task adversarial network based on an encoder-discriminator-generator architecture. The encoder extracts a disentangled feature representation for the factors of interest. The discriminators classify each of the factors as individual tasks. The encoder and the discriminators are trained cooperatively on factors of interest, but in an adversarial way on factors of distraction. The generator provides further regularization on the learned feature by reconstructing images with shared factors as the input image. We design a new optimization scheme to stabilize the adversarial optimization process when multiple distributions need to be aligned. The experiments on face recognition and font recognition tasks show that our method outperforms the state-of-the-art methods in terms of both recognizing the factors of interest and generalization to images with unseen variations.

****

Learning From Synthetic Data: Addressing Domain Shift for Semantic Segmentation
Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, Rama Chellappa; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3752-3761
Visual Domain Adaptation is a problem of immense importance in computer vision. Previous approaches showcase the inability of even deep neural networks to learn informative representations across domain shift. This problem is more severe for tasks where acquiring hand labeled data is extremely hard and tedious. In this work, we focus on adapting the representations learned by segmentation networks across synthetic and real domains. Contrary to previous approaches that use a simple adversarial objective or superpixel information to aid the process, we propose an approach based on Generative Adversarial Networks (GANs) that brings the embeddings closer in the learned feature space. To showcase the generality and scalability of our approach, we show that we can achieve state of the art results on two challenging scenarios of synthetic to real domain adaptation. Additional exploratory experiments show that our approach: (1) generalizes to unseen doma

ins and (2) results in improved alignment of source and target distributions.

****

Empirical Study of the Topology and Geometry of Deep Networks

Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard, Stefano Soatto; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3762-3770

The goal of this paper is to analyze the geometric properties of deep neural network image classifiers in the input space. We specifically study the topology of classification regions created by deep networks, as well as their associated decision boundary. Through a systematic empirical study, we show that state-of-the-art deep nets learn connected classification regions, and that the decision boundary in the vicinity of datapoints is flat along most directions. We further draw an essential connection between two seemingly unrelated properties of deep networks: their sensitivity to additive perturbations of the inputs, and the curvature of their decision boundary. The directions where the decision boundary is curved in fact characterize the directions to which the classifier is the most vulnerable. We finally leverage a fundamental asymmetry in the curvature of the decision boundary of deep nets, and propose a method to discriminate between original images, and images perturbed with small adversarial examples. We show the effectiveness of this purely geometric approach for detecting small adversarial perturbations in images, and for recovering the labels of perturbed images.

****

Boosting Domain Adaptation by Discovering Latent Domains

Massimiliano Mancini, Lorenzo Porzi, Samuel Rota Bulò, Barbara Caputo, Elisa Ricci; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3771-3780

Current Domain Adaptation (DA) methods based on deep architectures assume that the source samples arise from a single distribution. However, in practice most datasets can be regarded as mixtures of multiple domains. In these cases exploiting single-source DA methods for learning target classifiers may lead to sub-optimal, if not poor, results. In addition, in many applications it is difficult to manually provide the domain labels for all source data points, i.e. latent domains should be automatically discovered. This paper introduces a novel Convolutional Neural Network (CNN) architecture which (i) automatically discovers latent domains in visual datasets and (ii) exploits this information to learn robust target classifiers. Our approach is based on the introduction of two main components, which can be embedded into any existing CNN architecture: (i) a side branch that automatically computes the assignment of a source sample to a latent domain and (ii) novel layers that exploit domain membership information to appropriately align the distribution of the CNN internal feature representations to a reference distribution. We test our approach on publicly-available datasets, showing that it outperforms state-of-the-art multi-source DA methods by a large margin.

****

Shape From Shading Through Shape Evolution

Dawei Yang, Jia Deng; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3781-3790

In this paper, we address the shape-from-shading problem by training deep networks with synthetic images. Unlike conventional approaches that combine deep learning and synthetic imagery, we propose an approach that does not need any external shape dataset to render synthetic images. Our approach consists of two synergistic processes: the evolution of complex shapes from simple primitives, and the training of a deep network for shape-from-shading. The evolution generates better shapes guided by the network training, while the training improves by using the evolved shapes. We show that our approach achieves state-of-the-art performance on a shape-from-shading benchmark.

****

Weakly Supervised Instance Segmentation Using Class Peak Response

Yanzhao Zhou, Yi Zhu, Qixiang Ye, Qiang Qiu, Jianbin Jiao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3791-3800

Weakly supervised instance segmentation with image-level labels, instead of expensive pixel-level masks, remains unexplored. In this paper, we tackle this challenging problem by exploiting class peak responses to enable a classification network for instance mask extraction. With image labels supervision only, CNN classifiers in a fully convolutional manner can produce class response maps, which specify classification confidence at each image location. We observed that local maximums, i.e., peaks, in a class response map typically correspond to strong visual cues residing inside each instance. Motivated by this, we first design a process to stimulate peaks to emerge from a class response map. The emerged peaks are then back-propagated and effectively mapped to highly informative regions of each object instance, such as instance boundaries. We refer to the above maps generated from class peak responses as Peak Response Maps (PRMs). PRMs provide a fine-detailed instance-level representation, which allows instance masks to be extracted even with some off-the-shelf methods. To the best of our knowledge, we for the first time report results for the challenging image-level supervised instance segmentation task. Extensive experiments show that our method also boosts weakly supervised pointwise localization as well as semantic segmentation performance, and reports state-of-the-art results on popular benchmarks, including PASCAL VOC 2012 and MS COCO.
****

Collaborative and Adversarial Network for Unsupervised Domain Adaptation
Weichen Zhang, Wanli Ouyang, Wen Li, Dong Xu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3801-3809
In this paper, we propose a new unsupervised domain adaptation approach called Collaborative and Adversarial Network (CAN) through domain-collaborative and domain-adversarial training of neural networks. We use several domain classifiers on multiple CNN feature extraction layers/blocks, in which each domain classifier is connected to the hidden representations from one block and one loss function is defined based on the hidden presentation and the domain labels (e.g., source and target). We design a new loss function by integrating the losses from all blocks in order to learn informative representations from lower layers through collaborative learning and learn uninformative representations from higher layers through adversarial learning. We further extend our CAN method as Incremental CAN (iCAN), in which we iteratively select a set of pseudo-labelled target samples based on the image classifier and the last domain classifier from the previous training epoch and re-train our CAN model using the enlarged training set. Comprehensive experiments on two benchmark datasets Office and ImageCLEF-DA clearly demonstrate the effectiveness of our newly proposed approaches CAN and iCAN for unsupervised domain adaptation.
****

Environment Upgrade Reinforcement Learning for Non-Differentiable Multi-Stage Pipelines
Shuqin Xie, Zitian Chen, Chao Xu, Cewu Lu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3810-3819
Recent advances in multi-stage algorithms have shown great promise, but two important problems still remain. First of all, at inference time, information can't feed back from downstream to upstream. Second, at training time, end-to-end training is not possible if the overall pipeline involves non-differentiable functions, and so different stages can't be jointly optimized. In this paper, we propose a novel environment upgrade reinforcement learning framework to solve the feedback and joint optimization problems. Our framework re-links the downstream stage to the upstream stage by a reinforcement learning agent. While training the agent to improve final performance by refining the upstream stage's output, we also upgrade the downstream stage (environment) according to the agent's policy. In this way, agent policy and environment are jointly optimized. We propose a training algorithm for this framework to address the different training demands of agent and environment. Experiments on instance segmentation and human pose estimation demonstrate the effectiveness of the proposed framework.
****

Teaching Categories to Human Learners With Visual Explanations

Oisin Mac Aodha, Shihan Su, Yuxin Chen, Pietro Perona, Yisong Yue; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3820-3828

We study the problem of computer-assisted teaching with explanations. Conventional approaches for machine teaching typically only provide feedback at the instance level e.g., the category or label of the instance. However, it is intuitive that clear explanations from a knowledgeable teacher can significantly improve a student's ability to learn a new concept. To address these existing limitations, we propose a teaching framework that provides interpretable explanations as feedback and models how the learner incorporates this additional information. In the case of images, we show that we can automatically generate explanations that highlight the parts of the image that are responsible for the class label. Experiments on human learners illustrate that, on average, participants achieve better test set performance on challenging categorization tasks when taught with our interpretable approach compared to existing methods.

****

Density Adaptive Point Set Registration

Felix Järemo Lawin, Martin Danelljan, Fahad Shahbaz Khan, Per-Erik Forssén, Michael Felsberg; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3829-3837

Probabilistic methods for point set registration have demonstrated competitive results in recent years. These techniques estimate a probability distribution model of the point clouds. While such a representation has shown promise, it is highly sensitive to variations in the density of 3D points. This fundamental problem is primarily caused by changes in the sensor location across point sets. We revisit the foundations of the probabilistic registration paradigm. Contrary to previous works, we model the underlying structure of the scene as a latent probability distribution, and thereby induce invariance to point set density changes. Both the probabilistic model of the scene and the registration parameters are inferred by minimizing the Kullback-Leibler divergence in an Expectation Maximization based framework. Our density-adaptive registration successfully handles severe density variations commonly encountered in terrestrial Lidar applications. We perform extensive experiments on several challenging real-world Lidar datasets. The results demonstrate that our approach outperforms state-of-the-art probabilistic methods for multi-view registration, without the need of re-sampling.

****

Left-Right Comparative Recurrent Model for Stereo Matching

Zequn Jie, Pengfei Wang, Yonggen Ling, Bo Zhao, Yunchao Wei, Jiashi Feng, Wei Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3838-3846

Leveraging the disparity information from both left and right views is crucial for stereo disparity estimation. Left-right consistency check is an effective way to enhance the disparity estimation by referring to the information from the opposite view. However, the conventional left-right consistency check is an isolated post-processing step and heavily hand-crafted. This paper proposes a novel left-right comparative recurrent model to perform left-right consistency checking jointly with disparity estimation. At each recurrent step, the model produces disparity results for both views, and then performs online left-right comparison to identify the mismatched regions which may probably contain erroneously labeled pixels. A soft attention mechanism is introduced, which employs the learned error maps for better guiding the model to selectively focus on refining the unreliable regions at the next recurrent step. In this way, the generated disparity maps are progressively improved by the proposed recurrent model. Extensive evaluations on KITTI 2015, Scene Flow and Middlebury benchmarks validate the effectiveness of our model, demonstrating that state-of-the-art stereo disparity estimation results can be achieved by this new model.

****

Im2Pano3D: Extrapolating 360° Structure and Semantics Beyond the Field of View

Shuran Song, Andy Zeng, Angel X. Chang, Manolis Savva, Silvio Savarese, Thomas Funkhouser; Proceedings of the IEEE Conference on Computer Vision and Pattern Rec ognition (CVPR), 2018, pp. 3847-3856

We present Im2Pano3D, a convolutional neural network that generates a dense prediction of 3D structure and a probability distribution of semantic labels for a full 360 panoramic view of an indoor scene when given only a partial observation (<=50%) in the form of an RGB-D image. To make this possible, Im2Pano3D leverages strong contextual priors learned from large-scale synthetic and real-world indoor scenes. To ease the prediction of 3D structure, we propose to parameterize 3D surfaces with their plane equations and train the model to predict these parameters directly. To provide meaningful training supervision, we make use of multiple loss functions that consider both pixel level accuracy and global context consistency. Experiments demonstrate that Im2Pano3D is able to predict the semantics and 3D structure of the unobserved scene with more than 56% pixel accuracy and less than 0.52m average distance error, which is significantly better than alternative approaches.

****

Polarimetric Dense Monocular SLAM

Luwei Yang, Feitong Tan, Ao Li, Zhaopeng Cui, Yasutaka Furukawa, Ping Tan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3857-3866

This paper presents a novel polarimetric dense monocular SLAM (PDMS) algorithm based on a polarization camera. The algorithm exploits both photometric and polarimetric light information to produce more accurate and complete geometry. The polarimetric information allows us to recover the azimuth angle of surface normals from each video frame to facilitate dense reconstruction, especially at textureless or specular regions. There are two challenges in our approach: 1) surface azimuth angles from the polarization camera are very noisy; and 2) we need a near real-time solution for SLAM. Previous successful methods on polarimetric multi-view stereo are offline and require manually pre-segmented object masks to suppress the effects of erroneous angle information along boundaries. Our fully automatic approach efficiently iterates azimuth-based depth propagations, two-view depth consistency check, and depth optimization to produce a depthmap in real-time, where all the algorithmic steps are carefully designed to enable a GPU implementation. To our knowledge, this paper is the first to propose a photometric and polarimetric method for dense SLAM. We have qualitatively and quantitatively evaluated our algorithm against a few of competing methods, demonstrating the superior performance on various indoor and outdoor scenes.

****

A Unifying Contrast Maximization Framework for Event Cameras, With Applications to Motion, Depth, and Optical Flow Estimation

Guillermo Gallego, Henri Rebecq, Davide Scaramuzza; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3867-3876

We present a unifying framework to solve several computer vision problems with event cameras: motion, depth and optical flow estimation. The main idea of our framework is to find the point trajectories on the image plane that are best aligned with the event data by maximizing an objective function: the contrast of an image of warped events. Our method implicitly handles data association between the events, and therefore, does not rely on additional appearance information about the scene. In addition to accurately recovering the motion parameters of the problem, our framework produces motion-corrected edge-like images with high dynamic range that can be used for further scene analysis. The proposed method is not only simple, but more importantly, it is, to the best of our knowledge, the first method that can be successfully applied to such a diverse set of important vision tasks with event cameras.

****

Modeling Facial Geometry Using Compositional VAEs

Timur Bagautdinov, Chenglei Wu, Jason Saragih, Pascal Fua, Yaser Sheikh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3877-3886

We propose a method for learning non-linear face geometry representations using deep generative models. Our model is a variational autoencoder with multiple l evels of hidden variables where lower layers capture global geometry and higher ones encode more local deformations. Based on that, we propose a new parameterization of facial geometry that naturally decomposes the structure of the human face into a set of semantically meaningful levels of detail. This parameterization enables us to do model fitting while capturing varying level of detail under different types of geometrical constraints.

****

Tangent Convolutions for Dense Prediction in 3D

Maxim Tatarchenko, Jaesik Park, Vladlen Koltun, Qian-Yi Zhou; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3887-3896

We present an approach to semantic scene analysis using deep convolutional networks. Our approach is based on tangent convolutions - a new construction for convolutional networks on 3D data. In contrast to volumetric approaches, our method operates directly on surface geometry. Crucially, the construction is applicable to unstructured point clouds and other noisy real-world data. We show that tangent convolutions can be evaluated efficiently on large-scale point clouds with millions of points. Using tangent convolutions, we design a deep fully-convolutional network for semantic segmentation of 3D point clouds, and apply it to challenging real-world datasets of indoor and outdoor 3D environments. Experimental results show that the presented approach outperforms other recent deep network constructions in detailed analysis of large 3D scenes.

****

RayNet: Learning Volumetric 3D Reconstruction With Ray Potentials

Despoina Paschalidou, Osman Ulusoy, Carolin Schmitt, Luc Van Gool, Andreas Geiger; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3897-3906

In this paper, we consider the problem of reconstructing a dense 3D model using images captured from different views. Recent methods based on convolutional neural networks (CNN) allow learning the entire task from data. However, they do not incorporate the physics of image formation such as perspective geometry and occlusion. Instead, classical approaches based on Markov Random Fields (MRF) with ray-potentials explicitly model these physical processes, but they cannot cope with large surface appearance variations across different viewpoints. In this paper, we propose RayNet, which combines the strengths of both frameworks. RayNet integrates a CNN that learns view-invariant feature representations with an MRF that explicitly encodes the physics of perspective projection and occlusion. We train RayNet end-to-end using empirical risk minimization. We thoroughly evaluate our approach on challenging real-world datasets and demonstrate its benefits over a piece-wise trained baseline, hand-crafted models as well as other learning-based approaches.

****

Neural 3D Mesh Renderer

Hiroharu Kato, Yoshitaka Ushiku, Tatsuya Harada; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3907-3916

For modeling the 3D world behind 2D images, which 3D representation is most appropriate? A polygon mesh is a promising candidate for its compactness and geometric properties. However, it is not straightforward to model a polygon mesh from 2D images using neural networks because the conversion from a mesh to an image, or rendering, involves a discrete operation called rasterization, which prevents back-propagation. Therefore, in this work, we propose an approximate gradient for rasterization that enables the integration of rendering into neural networks. Using this renderer, we perform single-image 3D mesh reconstruction with silhouette image supervision and our system outperforms the existing voxel-based approach. Additionally, we perform gradient-based 3D mesh editing operations, such as 2D-to-3D style transfer and 3D DeepDream, with 2D supervision for the first time. These applications demonstrate the potential of the integration of a mesh renderer into neural networks and the effectiveness of our proposed renderer.

****

Structured Attention Guided Convolutional Neural Fields for Monocular Depth Esti mation

Dan Xu, Wei Wang, Hao Tang, Hong Liu, Nicu Sebe, Elisa Ricci; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3917-3925

Recent works have shown the benefit of integrating Conditional Random Fields (CRFs) models into deep architectures for improving pixel-level prediction tasks. Following this line of research, in this paper we introduce a novel approach for monocular depth estimation. Similarly to previous works, our method employs a continuous CRF to fuse multi-scale information derived from different layers of a front-end Convolutional Neural Network (CNN). Differently from past works, our approach benefits from a structured attention model which automatically regulates the amount of information transferred between corresponding features at different scales. Importantly, the proposed attention model is seamlessly integrated into the CRF, allowing end-to-end training of the entire architecture. Our extensive experimental evaluation demonstrates the effectiveness of the proposed method which is competitive with previous methods on the KITTI benchmark and outperforms the state of the art on the NYU Depth V2 dataset.

****

Automatic 3D Indoor Scene Modeling From Single Panorama

Yang Yang, Shi Jin, Ruiyang Liu, Sing Bing Kang, Jingyi Yu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3926-3934

We describe a system that automatically extracts 3D geometry of an indoor scene from a single 2D panorama. Our system recovers the spatial layout by finding the floor, walls, and ceiling; it also recovers shapes of typical indoor objects such as furniture. Using sampled perspective sub-views, we extract geometric cues (lines, vanishing points, orientation map, and surface normals) and semantic cues (saliency and object detection information). These cues are used for ground plane estimation and occlusion reasoning. The global spatial layout is inferred through a constraint graph on line segments and planar superpixels. The recovered layout is then used to guide shape estimation of the remaining objects using their normal information. Experiments on synthetic and real datasets show that our approach is state-of-the-art in both accuracy and efficiency. Our system can handle cluttered scenes with complex geometry that are challenging to existing techniques.

****

Extreme 3D Face Reconstruction: Seeing Through Occlusions

Anh Tu■n Tr■n, Tal Hassner, Iacopo Masi, Eran Paz, Yuval Nirkin, Gérard Medioni; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3935-3944

Existing single view, 3D face reconstruction methods can produce beautifully detailed 3D results, but typically only for near frontal, unobstructed viewpoints. We describe a system designed to provide detailed 3D reconstructions of faces viewed under extreme conditions, out of plane rotations, and occlusions. Motivated by the concept of bump mapping, we propose a layered approach which decouples estimation of a global shape from its mid-level details (e.g., wrinkles). We estimate a coarse 3D face shape which acts as a foundation and then separately layer this foundation with details represented by a bump map. We show how a deep convolutional encoder-decoder can be used to estimate such bump maps. We further show how this approach naturally extends to generate plausible details for occluded facial regions. We test our approach and its components extensively, quantitatively demonstrating the invariance of our estimated facial details. We further provide numerous qualitative examples showing that our method produces detailed 3D face shapes in viewing conditions where existing state of the art often break down.

****

Beyond Grobner Bases: Basis Selection for Minimal Solvers

Viktor Larsson, Magnus Oskarsson, Kalle Astrom, Alge Wallis, Zuzana Kukelova, Tomas Pajdla; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3945-3954

Many computer vision applications require robust estimation of the underlying geometry, in terms of camera motion and 3D structure of the scene. These robust methods often rely on running minimal solvers in a RANSAC framework. In this paper we show how we can make polynomial solvers based on the action matrix method faster, by careful selection of the monomial bases. These monomial bases have traditionally been based on a Grobner basis for the polynomial ideal. Here we describe how we can enumerate all such bases in an efficient way. We also show that going beyond Grobner bases leads to more efficient solvers in many cases. We present a novel basis sampling scheme that we evaluate on a number of problems.

****

Lions and Tigers and Bears: Capturing Non-Rigid, 3D, Articulated Shape From Images

Silvia Zuffi, Angjoo Kanazawa, Michael J. Black; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3955-3963

Animals are widespread in nature and the analysis of their shape and motion is important in many fields and industries. Modeling 3D animal shape, however, is difficult because the 3D scanning methods used to capture human shape are not applicable to wild animals or natural settings. Consequently, we propose a method to capture the detailed 3D shape of animals from images alone. The articulated and deformable nature of animals makes this problem extremely challenging, particularly in unconstrained environments with moving and uncalibrated cameras. To make this possible, we use a strong prior model of articulated animal shape that we fit to the image data. We then deform the animal shape in a canonical reference pose such that it matches image evidence when articulated and projected into multiple images. Our method extracts significantly more 3D shape detail than previous methods and is able to model new species, including the shape of an extinct animal, using only a few video frames. Additionally, the projected 3D shapes are accurate enough to facilitate the extraction of a realistic texture map from multiple frames.

****

Deep Cocktail Network: Multi-Source Unsupervised Domain Adaptation With Category Shift

Ruijia Xu, Ziliang Chen, Wangmeng Zuo, Junjie Yan, Liang Lin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3964-3973

Most existing unsupervised domain adaptation (UDA) methods are based upon the assumption that source labeled data come from an identical underlying distribution. Whereas in practical scenario, labeled instances are typically collected from diverse sources. Moreover, those sources may not completely share their categories, which further brings a category shift challenge to multi-source (unsupervised) domain adaptation (MDA). In this paper, we propose a deep cocktail network (DCTN), to battle the domain and category shifts among multiple sources. Motivated by the theoretical results in cite{mansour2009domain}, the target distribution can be represented as the weighted combination of source distributions, and, the training of MDA via DCTN is then performed as two alternating steps: i) It deploys multi-way adversarial learning to minimize the discrepancy between the target and each of the multiple source domains, which also obtains the source-specific perplexity scores to denote the possibilities that a target sample belongs to different source domains. ii) The multi-source category classifiers are integrated with the perplexity scores to classify target sample, and the pseudo-labeled target samples together with source samples are utilized to update the multi-source category classifier and the representation module. We evaluate DCTN in three domain adaptation benchmarks, which clearly demonstrate the superiority of our framework.

****

DOTA: A Large-Scale Dataset for Object Detection in Aerial Images

Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, Liangpei Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3974-3983

Object detection is an important and challenging problem in computer vision. Alt hough the past decade has witnessed major advances in object detection in natural scenes, such successes have been slow to aerial imagery, not only because of the huge variation in the scale, orientation and shape of the object instances on the earth's surface, but also due to the scarcity of well-annotated datasets of objects in aerial scenes. To advance object detection research in Earth Vision, also known as Earth Observation and Remote Sensing, we introduce a large-scale Dataset for Object deTection in Aerial images (DOTA). To this end, we collect 2806 aerial images from different sensors and platforms. Each image is of the size about 4000-by-4000 pixels and contains objects exhibiting a wide variety of scales, orientations, and shapes. These DOTA images are then annotated by experts in aerial image interpretation using 15 common object categories. The fully annotated DOTA images contains 188,282 instances, each of which is labeled by an arbitrary (8 d.o.f.) quadrilateral. To build a baseline for object detection in Earth Vision, we evaluate state-of-the-art object detection algorithms on DOTA. Experiments demonstrate that DOTA well represents real Earth Vision applications and are quite challenging.

****

Finding Beans in Burgers: Deep Semantic-Visual Embedding With Localization

Martin Engilberge, Louis Chevallier, Patrick Pérez, Matthieu Cord; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3984-3993

Several works have proposed to learn a two-path neural network that maps images and texts, respectively, to a same shared Euclidean space where geometry captures useful semantic relationships. Such a multi-modal embedding can be trained and used for various tasks, notably image captioning. In the present work, we introduce a new architecture of this type, with a visual path that leverages recent space-aware pooling mechanisms. Combined with a textual path which is jointly trained from scratch, our semantic-visual embedding offers a versatile model. Once trained under the supervision of captioned images, it yields new state-of-the-art performance on cross-modal retrieval. It also allows the localization of new concepts from the embedding space into any input image, delivering state-of-the-art result on the visual grounding of phrases.

****

Feature Super-Resolution: Make Machine See More Clearly

Weimin Tan, Bo Yan, Bahetiyaer Bare; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3994-4002

Identifying small size images or small objects is a notoriously challenging problem, as discriminative representations are difficult to learn from the limited information contained in them with poor-quality appearance and unclear object structure. Existing research works usually increase the resolution of low-resolution image in the pixel space in order to provide better visual quality for human viewing. However, the improved performance of such methods is usually limited or even trivial in the case of very small image size (we will show it in this paper explicitly). In this paper, different from image super-resolution (ISR), we propose a novel super-resolution technique called feature super-resolution (FSR), which aims at enhancing the discriminatory power of small size image in order to provide high recognition precision for machine. To achieve this goal, we propose a new Feature Super-Resolution Generative Adversarial Network (FSR-GAN) model that transforms the raw poor features of small size images to highly discriminative ones by performing super-resolution in the feature space. Our FSR-GAN consists of two subnetworks: a feature generator network G and a feature discriminator network D. By training the G and the D networks in an alternative manner, we encourage the G network to discover the latent distribution correlations between small size and large size images and then use G to improve the representations of small images. Extensive experiment results on Oxford5K, Paris, Holidays, and Flick100k datasets demonstrate that the proposed FSR approach can effectively enhance the discriminatory ability of features. Even when the resolution of query images is reduced greatly, e.g., 1/64 original size, the query feature enhanced by our FSR approach achieves surprisingly high retrieval performance at different image resolutions and increases the retrieval precision by 25% compared to the r aw query feature.

****

ClusterNet: Detecting Small Objects in Large Scenes by Exploiting Spatio-Temporal Information

Rodney LaLonde, Dong Zhang, Mubarak Shah; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4003-4012

Object detection in wide area motion imagery (WAMI) has drawn the attention of the computer vision research community for a number of years. WAMI proposes a number of unique challenges including extremely small object sizes, both sparse and densely-packed objects, and extremely large search spaces (large video frames). Nearly all state-of-the-art methods in WAMI object detection report that appearance-based classifiers fail in this challenging data and instead rely almost entirely on motion information in the form of background subtraction or frame-differencing. In this work, we experimentally verify the failure of appearance-based classifiers in WAMI, such as Faster R-CNN and a heatmap-based fully convolutional neural network (CNN), and propose a novel two-stage spatio-temporal CNN which effectively and efficiently combines both appearance and motion information to significantly surpass the state-of-the-art in WAMI object detection. To reduce the large search space, the first stage (ClusterNet) takes in a set of extremely large video frames, combines the motion and appearance information within the convolutional architecture, and proposes regions of objects of interest (ROOBI). These ROOBI can contain from one to clusters of several hundred objects due to the large video frame size and varying object density in WAMI. The second stage (FoveaNet) then estimates the centroid location of all objects in that given ROOBI simultaneously via heatmap estimation. The proposed method exceeds state-of-the-art results on the WPAFB 2009 dataset by 5-16% for moving objects and nearly 50% for stopped objects, as well as being the first proposed method in wide area motion imagery to detect completely stationary objects.

****

MaskLab: Instance Segmentation by Refining Object Detection With Semantic and Direction Features

Liang-Chieh Chen, Alexander Hermans, George Papandreou, Florian Schroff, Peng Wang, Hartwig Adam; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4013-4022

In this work, we tackle the problem of instance segmentation, the task of simultaneously solving object detection and semantic segmentation. Towards this goal, we present a model, called MaskLab, which produces three outputs: box detection, semantic segmentation, and direction prediction. Building on top of the Faster-RCNN object detector, the predicted boxes provide accurate localization of object instances. Within each region of interest, MaskLab performs foreground/background segmentation by combining semantic and direction prediction. Semantic segmentation assists the model in distinguishing between objects of different semantic classes including background, while the direction prediction, estimating each pixel's direction towards its corresponding center, allows separating instances of the same semantic class. Moreover, we explore the effect of incorporating recent successful methods from both segmentation and detection (eg, atrous convolution and hypercolumn). Our proposed model is evaluated on the COCO instance segmentation benchmark and shows comparable performance with other state-of-art models.

****

Hashing as Tie-Aware Learning to Rank

Kun He, Fatih Cakir, Sarah Adel Bargal, Stan Sclaroff; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4023-4032

Hashing, or learning binary embeddings of data, is frequently used in nearest neighbor retrieval. In this paper, we develop learning to rank formulations for hashing, aimed at directly optimizing ranking-based evaluation metrics such as Average Precision (AP) and Normalized Discounted Cumulative Gain (NDCG). We first observe that the integer-valued Hamming distance often leads to tied rankings, and propose to use tie-aware versions of AP and NDCG to evaluate hashing for retrieval. Then, to optimize tie-aware ranking metrics, we derive their continuous re

laxations, and perform gradient-based optimization with deep neural networks. Our results establish the new state-of-the-art for image retrieval by Hamming ranking in common benchmarks.
****

Classification-Driven Dynamic Image Enhancement
Vivek Sharma, Ali Diba, Davy Neven, Michael S. Brown, Luc Van Gool, Rainer Stiefelhagen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4033-4041
Convolutional neural networks rely on image texture and structure to serve as discriminative features to classify the image content. Image enhancement techniques can be used as preprocessing steps to help improve the overall image quality and in turn improve the overall effectiveness of a CNN. Existing image enhancement methods, however, are designed to improve the perceptual quality of an image for a human observer. In this paper, we are interested in learning CNNs that can emulate image enhancement and restoration, but with the overall goal to improve image classification and not necessarily human perception. To this end, we present a unified CNN architecture that uses a range of enhancement filters that can enhance image-specific details via end-to-end dynamic filter learning. We demonstrate the effectiveness of this strategy on four challenging benchmark datasets for fine-grained, object, scene and texture classification: CUB-200-2011, PASCAL-VOC2007, MIT-Indoor, and DTD. Experiments using our proposed enhancement shows promising results on all the datasets. In addition, our approach is capable of improving the performance of all generic CNN architectures.
****

Knowledge Aided Consistency for Weakly Supervised Phrase Grounding
Kan Chen, Jiyang Gao, Ram Nevatia; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4042-4050
Given a natural language query, a phrase grounding system aims to localize mentioned objects in an image. In weakly supervised scenario, mapping between image regions (i.e., proposals) and language is not available in the training set. Previous methods address this deficiency by training a grounding system via learning to reconstruct language information contained in input queries from predicted proposals. However, the optimization is solely guided by the reconstruction loss from the language modality, and ignores rich visual information contained in proposals and useful cues from external knowledge. In this paper, we explore the consistency contained in both visual and language modalities, and leverage complementary external knowledge to facilitate weakly supervised grounding. We propose a novel Knowledge Aided Consistency Network (KAC Net) which is optimized by reconstructing input query and proposal's information. To leverage complementary knowledge contained in the visual features, we introduce a Knowledge Based Pooling (KBP) gate to focus on query-related proposals. Experiments show that KAC Net provides a significant improvement on two popular datasets.
****

Who Let the Dogs Out? Modeling Dog Behavior From Visual Data
Kiana Ehsani, Hessam Bagherinezhad, Joseph Redmon, Roozbeh Mottaghi, Ali Farhadi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4051-4060
We introduce the task of directly modeling a visually intelligent agent. Computer vision typically focuses on solving various subtasks related to visual intelligence. We depart from this standard approach to computer vision; instead we directly model a visually intelligent agent. Our model takes visual information as input and directly predicts the actions of the agent. Toward this end we introduce DECADE, a large-scale dataset of ego-centric videos from a dog's perspective as well as her corresponding movements. Using this data we model how the dog acts and how the dog plans her movements. We show under a variety of metrics that given just visual input we can successfully model this intelligent agent in many situations. Moreover, the representation learned by our model encodes distinct information compared to representations trained on image classification, and our learned representation can generalize to other domains. In particular, we show strong results on the task of walkable surface estimation by using this dog modeli

ng task as representation learning.
****
Pseudo Mask Augmented Object Detection
Xiangyun Zhao, Shuang Liang, Yichen Wei; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4061-4070
In this work, we present a novel and effective framework to facilitate object detection with the instance-level segmentation information that is only supervised by bounding box annotation. Starting from the joint object detection and instance segmentation network, we propose to recursively estimate the pseudo ground-truth object masks from the instance-level object segmentation network training, and then enhance the detection network with top-down segmentation feedbacks. The pseudo ground truth mask and network parameters are optimized alternatively to mutually benefit each other. To obtain the promising pseudo masks in each iteration, we embed a graphical inference that incorporates the low-level image appearance consistency and the bounding box annotations to refine the segmentation masks predicted by the segmentation network. Our approach progressively improves the object detection performance by incorporating the detailed pixel-wise information learned from the weakly-supervised segmentation network. Extensive evaluation on the detection task in PASCAL VOC 2007 and 2012 verifies that the proposed approach is effective.
****
Dual Skipping Networks
Changmao Cheng, Yanwei Fu, Yu-Gang Jiang, Wei Liu, Wenlian Lu, Jianfeng Feng, Xiangyang Xue; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4071-4079
Inspired by the recent neuroscience studies on the left-right asymmetry of the human brain in processing low and high spatial frequency information, this paper introduces a dual skipping network which carries out coarse-to-fine object categorization. Such a network has two branches to simultaneously deal with both coarse and fine-grained classification tasks. Specifically, we propose a layer-skipping mechanism that learns a gating network to predict which layers to skip in the testing stage. This layer-skipping mechanism endows the network with good flexibility and capability in practice. Evaluations are conducted on several widely used coarse-to-fine object categorization benchmarks, and promising results are achieved by our proposed network model.
****
Memory Matching Networks for One-Shot Image Recognition
Qi Cai, Yingwei Pan, Ting Yao, Chenggang Yan, Tao Mei; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4080-4088
In this paper, we introduce the new ideas of augmenting Convolutional Neural Networks (CNNs) with Memory and learning to learn the network parameters for the unlabelled images on the fly in one-shot learning. Specifically, we present Memory Matching Networks (MM-Net) --- a novel deep architecture that explores the training procedure, following the philosophy that training and test conditions must match. Technically, MM-Net writes the features of a set of labelled images (support set) into memory and reads from memory when performing inference to holistically leverage the knowledge in the set. Meanwhile, a Contextual Learner employs the memory slots in a sequential manner to predict the parameters of CNNs for unlabelled images. The whole architecture is trained by once showing only a few examples per class and switching the learning from minibatch to minibatch, which is tailored for one-shot learning when presented with a few examples of new categories at test time. Unlike the conventional one-shot learning approaches, our MM-Net could output one unified model irrespective of the number of shots and categories. Extensive experiments are conducted on two public datasets, i.e., Omniglot and emph{mini}ImageNet, and superior results are reported when compared to state-of-the-art approaches. More remarkably, our MM-Net improves one-shot accuracy on Omniglot from 98.95% to 99.28% and from 49.21% to 53.37% on emph{mini}ImageNet.
****
IQA: Visual Question Answering in Interactive Environments

Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, Ali Farhadi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4089-4098

We introduce Interactive Question Answering (IQA), the task of answering questions that require an autonomous agent to interact with a dynamic visual environment. IQA presents the agent with a scene and a question, like: “Are there any apples in the fridge?” The agent must navigate around the scene, acquire visual understanding of scene elements, interact with objects (e.g. open refrigerators) and plan for a series of actions conditioned on the question. Popular reinforcement learning approaches with a single controller perform poorly on IQA owing to the large and diverse state space. We propose the Hierarchical Interactive Memory Network (HIMN), consisting of a factorized set of controllers, allowing the system to operate at multiple levels of temporal abstraction. To evaluate HIMN, we introduce IQUAD V1, a new dataset built upon AI2-THOR [35], a simulated photo-realistic environment of configurable indoor scenes with interactive objects. IQUAD V1 has 75,000 questions, each paired with a unique scene configuration. Our experiments show that our proposed model outperforms popular single controller based methods on IQUAD V1. For sample questions and results, please view our video: https://youtu.be/pXd3C-1jr98.

****

Pose Transferrable Person Re-Identification

Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, Jianguo Hu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4099-4108

Person re-identification (ReID) is an important task in the field of intelligent security. A key challenge is how to capture human pose variations, while existing benchmarks (i.e., Market1501, DukeMTMC-reID, CUHK03, etc.) do NOT provide sufficient pose coverage to train a robust ReID system. To address this issue, we propose a pose-transferrable person ReID framework which utilizes pose-transferred sample augmentations (i.e., with ID supervision) to enhance ReID model training. On one hand, novel training samples with rich pose variations are generated via transferring pose instances from MARS dataset, and they are added into the target dataset to facilitate robust training. On the other hand, in addition to the conventional discriminator of GAN (i.e., to distinguish between REAL/FAKE samples), we propose a novel guider sub-network which encourages the generated sample (i.e., with novel pose) towards better satisfying the ReID loss (i.e., cross-entropy ReID loss, triplet ReID loss). In the meantime, an alternative optimization procedure is proposed to train the proposed Generator-Guider-Discriminator network. Experimental results on Market-1501, DukeMTMC-reID and CUHK03 show that our method achieves great performance improvement, and outperforms most state-of-the-art methods without elaborate designing the ReID model.

****

Large Scale Fine-Grained Categorization and Domain-Specific Transfer Learning

Yin Cui, Yang Song, Chen Sun, Andrew Howard, Serge Belongie; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4109-4118

Transferring the knowledge learned from large scale datasets (e.g., ImageNet) via fine-tuning offers an effective solution for domain-specific fine-grained visual categorization (FGVC) tasks (e.g., recognizing bird species or car make & model). In such scenarios, data annotation often calls for specialized domain knowledge and thus is difficult to scale. In this work, we first tackle a problem in large scale FGVC. Our method won first place in iNaturalist 2017 large scale species classification challenge. Central to the success of our approach is a training scheme that uses higher image resolution and deals with the long-tailed distribution of training data. Next, we study transfer learning via fine-tuning from large scale datasets to small scale, domain-specific FGVC datasets. We propose a measure to estimate domain similarity via Earth Mover's Distance and demonstrate that transfer learning benefits from pre-training on a source domain that is similar to the target domain by this measure. Our proposed transfer learning outperforms ImageNet pre-training and obtains state-of-the-art results on multiple

commonly used FGVC datasets.
****

Data Distillation: Towards Omni-Supervised Learning
Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, Kaiming He; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4119-4128
We investigate omni-supervised learning, a special regime of semi-supervised learning in which the learner exploits all available labeled data plus internet-scale sources of unlabeled data. Omni-supervised learning is lower-bounded by performance on existing labeled datasets, offering the potential to surpass state-of-the-art fully supervised methods. To exploit the omni-supervised setting, we propose data distillation, a method that ensembles predictions from multiple transformations of unlabeled data, using a single model, to automatically generate new training annotations. We argue that visual recognition models have recently become accurate enough that it is now possible to apply classic ideas about self-training to challenging real-world data. Our experimental results show that in the cases of human keypoint detection and general object detection, state-of-the-art models trained with data distillation surpass the performance of using labeled data from the COCO dataset alone.
****

Object Referring in Videos With Language and Human Gaze
Arun Balajee Vasudevan, Dengxin Dai, Luc Van Gool; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4129-4138
We investigate the problem of object referring (OR) i.e. to localize a target object in a visual scene coming with a language description. Humans perceive the world more as continued video snippets than as static images, and describe objects not only by their appearance, but also by their spatio-temporal context and motion features. Humans also gaze at the object when they issue a referring expression. Existing works for OR mostly focus on static images only, which fall short in providing many such cues. This paper addresses OR in videos with language and human gaze. To that end, we present a new video dataset for OR, with 30,000 objects over 5,000 stereo video sequences annotated for their descriptions and gaze. We further propose a novel network model for OR in videos, by integrating appearance, motion, gaze, and spatio-temporal context into one network. Experimental results show that our method effectively utilizes motion cues, human gaze, and spatio-temporal context. Our method outperforms previous OR methods.
****

Feature Selective Networks for Object Detection
Yao Zhai, Jingjing Fu, Yan Lu, Houqiang Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4139-4147
Objects for detection usually have distinct characteristics in different sub-regions and different aspect ratios. However, in prevalent two-stage object detection methods, Region-of-Interest (RoI) features are extracted by RoI pooling with little emphasis on these translation-variant feature components. We present feature selective networks to reform the feature representations of RoIs by exploiting their disparities among sub-regions and aspect ratios. Our network produces the sub-region attention bank and aspect ratio attention bank for the whole image. The RoI-based sub-region attention map and aspect ratio attention map are selectively pooled from the banks, and then used to refine the original RoI features for RoI classification. Equipped with a light-weight detection subnetwork, our network gets a consistent boost in detection performance based on general ConvNet backbones (ResNet-101, GoogLeNet and VGG-16). Without bells and whistles, our detectors equipped with ResNet-101 achieve more than 3% mAP improvement compared to counterparts on PASCAL VOC 2007, PASCAL VOC 2012 and MS COCO datasets.
****

Learning a Discriminative Filter Bank Within a CNN for Fine-Grained Recognition
Yaming Wang, Vlad I. Morariu, Larry S. Davis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4148-4157
Compared to earlier multistage frameworks using CNN features, recent end-to-end deep approaches for fine-grained recognition essentially enhance the mid-level l

earning capability of CNNs. Previous approaches achieve this by introducing an auxiliary network to infuse localization information into the main classification network, or a sophisticated feature encoding method to capture higher order feature statistics. We show that mid-level representation learning can be enhanced within the CNN framework, by learning a bank of convolutional filters that capture class-specific discriminative patches without extra part or bounding box annotations. Such a filter bank is well structured, properly initialized and discriminatively learned through a novel asymmetric multi-stream architecture with convolutional filter supervision and a non-random layer initialization. Experimental results show that our approach achieves state-of-the-art on three publicly available fine-grained recognition datasets (CUB-200-2011, Stanford Cars and FGVC-Aircraft). Ablation studies and visualizations are further provided to understand our approach.

****

Grounding Referring Expressions in Images by Variational Context

Hanwang Zhang, Yulei Niu, Shih-Fu Chang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4158-4166

We focus on grounding (i.e., localizing or linking) referring expressions in images, e.g., ``largest elephant standing behind baby elephant''.  This is a general yet challenging vision-language task since it does not only require the localization of objects, but also the multimodal comprehension of context --- visual attributes (e.g., ``largest'', ``baby'') and relationships (e.g., ``behind'') that help to distinguish the referent from other objects, especially those of the same category. Due to the exponential complexity involved in modeling the context associated with multiple image regions, existing work oversimplifies this task to pairwise region modeling by multiple instance learning. In this paper, we propose a variational Bayesian method, called Variational Context, to solve the problem of complex context modeling in referring expression grounding. Our model exploits the reciprocal relation between the referent and context, i.e., either of them influences the estimation of the posterior distribution of the other, and thereby the search space of context can be greatly reduced. We also extend the model to the unsupervised setting where no annotation for the referent is available. Extensive experiments on various benchmarks show consistent improvement over state-of-the-art methods in both supervised and unsupervised settings. The code is available at url{https://github.com/yuleiniu/vc/

****

Dynamic Graph Generation Network: Generating Relational Knowledge From Diagrams

Daesik Kim, YoungJoon Yoo, Jee-Soo Kim, SangKuk Lee, Nojun Kwak; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4167-4175

In this work, we introduce a new algorithm for analyzing a diagram, which contains visual and textual information in an abstract and integrated way. Whereas diagrams contain richer information compared with individual image-based or language-based data, proper solutions for automatically understanding them have not been proposed due to their innate characteristics of multi-modality and arbitrariness of layouts. To tackle this problem, we propose a unified diagram-parsing network for generating knowledge from diagrams based on an object detector and a recurrent neural network designed for a graphical structure. Specifically, we propose a dynamic graph-generation network that is based on dynamic memory and graph theory. We explore the dynamics of information in a diagram with activation of gates in gated recurrent unit (GRU) cells. On publicly available diagram datasets, our model demonstrates a state-of-the-art result that outperforms other baselines. Moreover, further experiments on question answering shows potentials of the proposed method for various applications.

****

A Network Architecture for Point Cloud Classification via Automatic Depth Images Generation

Riccardo Roveri, Lukas Rahmann, Cengiz Oztireli, Markus Gross; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4176-4184

We propose a novel neural network architecture for point cloud classification. Our key idea is to automatically transform the 3D unordered input data into a set of useful 2D depth images, and classify them by exploiting well performing image classification CNNs. We present new differentiable module designs to generate depth images from a point cloud. These modules can be combined with any network architecture for processing point clouds. We utilize them in combination with state-of-the-art classification networks, and get results competitive with the state of the art in point cloud classification. Furthermore, our architecture automatically produces informative images representing the input point cloud, which could be used for further applications such as point cloud visualization.

****

Towards Dense Object Tracking in a 2D Honeybee Hive

Katarzyna Bozek, Laetitia Hebert, Alexander S. Mikheyev, Greg J. Stephens; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4185-4193

From human crowds to cells in a tissue, the detection and efficient tracking of multiple objects in dense configurations is an important and unsolved problem. In the past, limitations of image analysis have restricted studies of dense groups to tracking one individual, a set of marked individuals, or to coarse-grained group-level dynamics, all of which yield incomplete information. Here, we combine the power of convolutional neural networks (CNNs) with the model environment of a honeybee hive to develop an automated method for the recognition of all individuals in a dense group based on raw image data. In the proposed solution, we create new, adapted individual labeling and use segmentation architecture U-Net with a specific loss function to predict both object location and orientation. We additionally leverage time series image data to exploit both structural and temporal regularities in the the tracked objects in a recurrent manner. This allowed us to achieve near human-level performance on real-world image data while dramatically reducing original network size to 6% of the initial parameters. Given the novel application of CNNs in this study, we generate extensive problem-specific image data in which labeled examples are produced through a custom interface with Amazon Mechanical Turk. This dataset contains over 375,000 labeled bee instances moving across 720 video frames with 2 fps sampling and represents an extensive resource for development and testing of dense object recognition and tracking methods. With our method we correctly detect 96% of individuals with a location error of ~7% of a typical body dimension, and orientation error of 12 degrees, approximating the variability in labeling by human raters with ~9% body dimension variation in position and 8 degrees orientation variation. Our study represents an important step towards efficient image-based dense object tracking by allowing for the accurate determination of object location and orientation across time-series image data efficiently within one network architecture.

****

Long-Term On-Board Prediction of People in Traffic Scenes Under Uncertainty

Apratim Bhattacharyya, Mario Fritz, Bernt Schiele; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4194-4202

Progress towards advanced systems for assisted and autonomous driving is leveraging recent advances in recognition and segmentation methods. Yet, we are still facing challenges in bringing reliable driving to inner cities, as those are composed of highly dynamic scenes observed from a moving platform at considerable speeds. Anticipation becomes a key element in order to react timely and prevent accidents. In this paper we argue that it is necessary to predict at least 1 second and we thus propose a new model that jointly predicts ego motion and people trajectories over such large time horizons. We pay particular attention to modeling the uncertainty of our estimates arising from the non-deterministic nature of natural traffic scenes. Our experimental results show that it is indeed possible to predict people trajectories at the desired time horizons and that our uncertainty estimates are informative of the prediction error. We also show that both sequence modeling of trajectories as well as our novel method of long term odometry prediction are essential for best performance.

****

Single-Shot Refinement Neural Network for Object Detection

Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, Stan Z. Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4203-4212

For object detection, the two-stage approach (e.g., Faster R-CNN) has been achieving the highest accuracy, whereas the one-stage approach (e.g., SSD) has the advantage of high efficiency. To inherit the merits of both while overcoming their disadvantages, in this paper, we propose a novel single-shot based detector, called RefineDet, that achieves better accuracy than two-stage methods and maintains comparable efficiency of one-stage methods. RefineDet consists of two inter-connected modules, namely, the anchor refinement module and the object detection module. Specifically, the former aims to (1) filter out negative anchors to reduce search space for the classifier, and (2) coarsely adjust the locations and sizes of anchors to provide better initialization for the subsequent regressor. The latter module takes the refined anchors as the input from the former to further improve the regression accuracy and predict multi-class label. Meanwhile, we design a transfer connection block to transfer the features in the anchor refinement module to predict locations, sizes and class labels of objects in the object detection module. The multi-task loss function enables us to train the whole network in an end-to-end way. Extensive experiments on PASCAL VOC 2007, PASCAL VOC 2012, and MS COCO demonstrate that RefineDet achieves state-of-the-art detection accuracy with high efficiency. Code is available at https://github.com/sfzhang15/RefineDet.

****

Video Captioning via Hierarchical Reinforcement Learning

Xin Wang, Wenhu Chen, Jiawei Wu, Yuan-Fang Wang, William Yang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4213-4222

Video captioning is the task of automatically generating a textual description of the actions in a video. Although previous work (e.g. sequence-to-sequence model) has shown promising results in abstracting a coarse description of a short video, it is still very challenging to caption a video containing multiple fine-grained actions with a detailed description. This paper aims to address the challenge by proposing a novel hierarchical reinforcement learning framework for video captioning, where a high-level Manager module learns to design sub-goals and a low-level Worker module recognizes the primitive actions to fulfill the sub-goal. With this compositional framework to reinforce video captioning at different levels, our approach significantly outperforms all the baseline methods on a newly introduced large-scale dataset for fine-grained video captioning. Furthermore, our non-ensemble model has already achieved the state-of-the-art results on the widely-used MSR-VTT dataset.

****

Tips and Tricks for Visual Question Answering: Learnings From the 2017 Challenge

Damien Teney, Peter Anderson, Xiaodong He, Anton van den Hengel; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4223-4232

This paper presents a state-of-the-art model for visual question answering (VQA), which won the first place in the 2017 VQA Challenge. VQA is a task of significant importance for research in artificial intelligence, given its multimodal nature, clear evaluation protocol, and potential real-world applications. The performance of deep neural networks for VQA is very dependent on choices of architectures and hyperparameters. To help further research in the area, we describe in detail our high-performing, though relatively simple model. Through a massive exploration of architectures and hyperparameters representing more than 3,000 GPU-hours, we identified tips and tricks that lead to its success, namely: sigmoid outputs, soft training targets, image features from bottom-up attention, gated tanh activations, output embeddings initialized using GloVe and Google Images, large mini-batches, and smart shuffling of training data. We provide a detailed analysis of their impact on performance to assist others in making an appropriate selection.

****

Learning to Segment Every Thing

Ronghang Hu, Piotr Dollár, Kaiming He, Trevor Darrell, Ross Girshick; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4233-4241

Most methods for object instance segmentation require all training examples to be labeled with segmentation masks. This requirement makes it expensive to annotate new categories and has restricted instance segmentation models to ~100 well-annotated classes. The goal of this paper is to propose a new partially supervised training paradigm, together with a novel weight transfer function, that enables training instance segmentation models on a large set of categories all of which have box annotations, but only a small fraction of which have mask annotations. These contributions allow us to train Mask R-CNN to detect and segment 3000 visual concepts using box annotations from the Visual Genome dataset and mask annotations from the 80 classes in the COCO dataset. We evaluate our approach in a controlled study on the COCO dataset. This work is a first step towards instance segmentation models that have broad comprehension of the visual world.

****

Self-Supervised Adversarial Hashing Networks for Cross-Modal Retrieval

Chao Li, Cheng Deng, Ning Li, Wei Liu, Xinbo Gao, Dacheng Tao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4242-4251

Thanks to the success of deep learning, cross-modal retrieval has made significant progress recently. However, there still remains a crucial bottleneck: how to bridge the modality gap to further enhance the retrieval accuracy. In this paper, we propose a self-supervised adversarial hashing (SSAH) approach, which lies among the early attempts to incorporate adversarial learning into cross-modal hashing in a self-supervised fashion. The primary contribution of this work is that two adversarial networks are leveraged to maximize the semantic correlation and consistency of the representations between different modalities. In addition, we harness a self-supervised semantic network to discover high-level semantic information in the form of multi-label annotations. Such information guides the feature learning process and preserves the modality relationships in both the common semantic space and the Hamming space. Extensive experiments carried out on three benchmark datasets validate that the proposed SSAH surpasses the state-of-the-art methods.

****

Parallel Attention: A Unified Framework for Visual Object Discovery Through Dialogs and Queries

Bohan Zhuang, Qi Wu, Chunhua Shen, Ian Reid, Anton van den Hengel; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4252-4261

Recognising objects according to a pre-defined fixed set of class labels has been well studied in the Computer Vision. There are a great many practical applications where the subjects that may be of interest are not known beforehand, or so easily delineated, however. In many of these cases natural language dialog is a natural way to specify the subject of interest, and the task achieving this capability (a.k.a, Referring Expression Comprehension) has recently attracted attention.To this end we propose a unified framework, the ParalleL AttentioN (PLAN) network, to discover the object in an image that is being referred to in variable length natural expression descriptions, from short phrases query to long multi-round dialogs. The PLAN network has two attention mechanisms that relate parts of the expressions to both the global visual content and also directly to object candidates. Furthermore, the attention mechanisms are recurrent, making the referring process visualizable and explainable. The attended information from these dual sources are combined to reason about the referred object. These two attention mechanisms can be trained in parallel and we find the combined system outperforms the state-of-art on several benchmarked datasets with different length language input, such as RefCOCO, RefCOCO+ and GuessWhat?!.

****

Zigzag Learning for Weakly Supervised Object Detection
Xiaopeng Zhang, Jiashi Feng, Hongkai Xiong, Qi Tian; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4262-4270
This paper addresses weakly supervised object detection with only image-level supervision at training stage. Previous approaches train detection models with entire images all at once, making the models prone to being trapped in sub-optimums due to the introduced false positive examples. Unlike them, we propose a zigzag learning strategy to simultaneously discover reliable object instances and prevent the model from overfitting initial seeds. Towards this goal, we first develop a criterion named mean Energy Accumulation Scores (mEAS) to automatically measure and rank localization difficulty of an image containing the target object, and accordingly learn the detector progressively by feeding examples with increasing difficulty. In this way, the model can be well prepared by training on easy examples for learning from more difficult ones and thus gain a stronger detection ability more efficiently. Furthermore, we introduce a novel masking regularization strategy over the high level convolutional feature maps to avoid overfitting initial samples. These two modules formulate a zigzag learning process, where progressive learning endeavors to discover reliable object instances, and masking regularization increases the difficulty of finding object instances properly. We achieve 47.6% mAP on PASCAL VOC 2007, surpassing the state-of-the-arts by a large margin.
****
Attentive Fashion Grammar Network for Fashion Landmark Detection and Clothing Category Classification
Wenguan Wang, Yuanlu Xu, Jianbing Shen, Song-Chun Zhu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4271-4280
This paper proposes a knowledge-guided fashion network to solve the problem of visual fashion analysis, e.g., fashion landmark localization and clothing category classification. The suggested fashion model is leveraged with high-level human knowledge in this domain. We propose two important fashion grammars: (i) dependency grammar capturing kinematics-like relation, and (ii) symmetry grammar accounting for the bilateral symmetry of clothes. We introduce Bidirectional Convolutional Recurrent Neural Networks (BCRNNs) for efficiently approaching message passing over grammar topologies, and producing regularized landmark layouts. For enhancing clothing category classification, our fashion network is encoded with two novel attention mechanisms, i.e., landmark-aware attention and category-driven attention. The former enforces our network to focus on the functional parts of clothes, and learns domain-knowledge centered representations, leading to a supervised attention mechanism. The latter is goal-driven, which directly enhances task-related features and can be learned in an implicit, top-down manner. Experimental results on large-scale fashion datasets demonstrate the superior performance of our fashion grammar network.
****
Generalized Zero-Shot Learning via Synthesized Examples
Vinay Kumar Verma, Gundeep Arora, Ashish Mishra, Piyush Rai; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4281-4289
We present a generative framework for generalized zero-shot learning where the training and test classes are not necessarily disjoint. Built upon a variational autoencoder based architecture, consisting of a probabilistic encoder and a probabilistic emph{conditional} decoder, our model can generate novel exemplars from seen/unseen classes, given their respective class attributes. These exemplars can subsequently be used to train any off-the-shelf classification model. One of the key aspects of our encoder-decoder architecture is a feedback-driven mechanism in which a discriminator (a multivariate regressor) learns to map the generated exemplars to the corresponding class attribute vectors, leading to an improved generator. Our model's ability to generate and leverage examples from unseen classes to train the classification model naturally helps to mitigate the bias towards predicting seen classes in generalized zero-shot learning settings. Through a comprehensive set of experiments, we show that our model outperforms several

state-of-the-art methods, on several benchmark datasets, for both standard as well as generalized zero-shot learning.

****

Partially Shared Multi-Task Convolutional Neural Network With Local Constraint for Face Attribute Learning

Jiajiong Cao, Yingming Li, Zhongfei Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4290-4299

In this paper, we study the face attribute learning problem by considering the identity information and attribute relationships simultaneously. In particular, we first introduce a Partially Shared Multi-task Convolutional Neural Network (PS-MCNN), in which four Task Specific Networks (TSNets) and one Shared Network (SNet) are connected by Partially Shared (PS) structures to learn better shared and task specific representations. To utilize identity information to further boost the performance, we introduce a local learning constraint which minimizes the difference between the representations of each sample and its local geometric neighbours with the same identity. Consequently, we present a local constraint regularized multi-task network, called Partially Shared Multi-task Convolutional Neural Network with Local Constraint (PS-MCNN-LC), where PS structure and local constraint are integrated together to help the framework learn better attribute representations. The experimental results on CelebA and LFWA demonstrate the promise of the proposed methods.

****

SYQ: Learning Symmetric Quantization for Efficient Deep Neural Networks

Julian Faraone, Nicholas Fraser, Michaela Blott, Philip H.W. Leong; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4300-4309

Inference for state-of-the-art deep neural networks is computationally expensive, making them difficult to deploy on constrained hardware environments. An efficient way to reduce this complexity is to quantize the weight parameters and/or activations during training by approximating their distributions with a limited entry codebook. For very low-precisions, such as binary or ternary networks with 1-8-bit activations, the information loss from quantization leads to significant accuracy degradation due to large gradient mismatches between the forward and backward functions. In this paper, we introduce a quantization method to reduce this loss by learning a symmetric codebook for particular weight subgroups. These subgroups are determined based on their locality in the weight matrix, such that the hardware simplicity of the low-precision representations is preserved. Empirically, we show that symmetric quantization can substantially improve accuracy for networks with extremely low-precision weights and activations. We also demonstrate that this representation imposes minimal or no hardware implications to more coarse-grained approaches. Source code is available at https://www.github.com/julianfaraone/SYQ.

****

DS*: Tighter Lifting-Free Convex Relaxations for Quadratic Matching Problems

Florian Bernard, Christian Theobalt, Michael Moeller; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4310-4319

In this work we study convex relaxations of quadratic optimisation problems over permutation matrices. While existing semidefinite programming approaches can achieve remarkably tight relaxations, they have the strong disadvantage that they lift the original n^2-dimensional variable to an n^4-dimensional variable, which limits their practical applicability. In contrast, here we present a lifting-free convex relaxation that is provably at least as tight as existing (lifting-free) convex relaxations. We demonstrate experimentally that our approach is superior to existing convex and non-convex methods for various problems, including image arrangement and multi-graph matching.

****

Deep Mutual Learning

Ying Zhang, Tao Xiang, Timothy M. Hospedales, Huchuan Lu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4320-4328

Model distillation is an effective and widely used technique to transfer knowledge from a teacher to a student network. The typical application is to transfer from a powerful large network or ensemble to a small network, in order to meet the low-memory or fast execution requirements. In this paper, we present a deep mutual learning (DML) strategy. Different from the one-way transfer between a static pre-defined teacher and a student in model distillation, with DML, an ensemble of students learn collaboratively and teach each other throughout the training process. Our experiments show that a variety of network architectures benefit from mutual learning and achieve compelling results on both category and instance recognition tasks. Surprisingly, it is revealed that no prior powerful teacher network is necessary -- mutual learning of a collection of simple student networks works, and moreover outperforms distillation from a more powerful yet static teacher.
****

Coupled End-to-End Transfer Learning With Generalized Fisher Information
Shixing Chen, Caojin Zhang, Ming Dong; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4329-4338
In transfer learning, one seeks to transfer related information from source tasks with sufficient data to help with the learning of target task with only limited data. In this paper, we propose a novel Coupled End-to-end Transfer Learning (CETL) framework, which mainly consists of two convolutional neural networks (source and target) that connect to a shared decoder. A novel loss function, the coupled loss, is used for CETL training. From a theoretical perspective, we demonstrate the rationale of the coupled loss by establishing a learning bound for CETL. Moreover, we introduce the generalized Fisher information to improve multi-task optimization in CETL. From a practical aspect, CETL provides a unified and highly flexible solution for various learning tasks such as domain adaption and knowledge distillation. Empirical result shows the superior performance of CETL on cross-domain and cross-task image classification.
****

Residual Parameter Transfer for Deep Domain Adaptation
Artem Rozantsev, Mathieu Salzmann, Pascal Fua; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4339-4348
The goal of Deep Domain Adaptation is to make it possible to use Deep Nets trained in one domain where there is enough annotated training data in another where there is little or none. Most current approaches have focused on learning feature representations that are invariant to the changes that occur when going from one domain to the other, which means using the same network parameters in both domains. While some recent algorithms explicitly model the changes by adapting the network parameters, they either severely restrict the possible domain changes, or significantly increase the number of model parameters. By contrast, we introduce a network architecture that includes auxiliary residual networks, which we train to predict the parameters in the domain with little annotated data from those in the other one. This architecture enables us to flexibly preserve the similarities between domains where they exist and model the differences when necessary. We demonstrate that our approach yields higher accuracy than state-of-the-art methods without undue complexity.
****

High-Order Tensor Regularization With Application to Attribute Ranking
Kwang In Kim, Juhyun Park, James Tompkin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4349-4357
When learning functions on manifolds, we can improve performance by regularizing with respect to the intrinsic manifold geometry rather than the ambient space. However, when regularizing tensor learning, calculating the derivatives along this intrinsic geometry is not possible, and so existing approaches are limited to regularizing in Euclidean space. Our new method for intrinsically regularizing and learning tensors on Riemannian manifolds introduces a surrogate object to encapsulate the geometric characteristic of the tensor. Regularizing this instead allows us to learn non-symmetric and high-order tensors. We apply our approach to the relative attributes problem, and we demonstrate that explicitly regularizi

ng high-order relationships between pairs of data points improves performance.
****

Learning to Localize Sound Source in Visual Scenes

Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, In So Kweon; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4358-4366

Visual events are usually accompanied by sounds in our daily lives. We pose the question: Can the machine learn the correspondence between visual scene and the sound, and localize the sound source only by observing sound and visual scene pairs like human? In this paper, we propose a novel unsupervised algorithm to address the problem of localizing the sound source in visual scenes. A two-stream network structure which handles each modality, with attention mechanism is developed for sound source localization. Moreover, although our network is formulated within the unsupervised learning framework, it can be extended to a unified architecture with a simple modification for the supervised and semi-supervised learning settings as well. Meanwhile, a new sound source dataset is developed for performance evaluation. Our empirical evaluation shows that the unsupervised method eventually go through false conclusion in some cases. We show that even with a few supervision, i.e., semi-supervised setup, false conclusion is able to be corrected effectively.
****

Dynamic Few-Shot Visual Learning Without Forgetting

Spyros Gidaris, Nikos Komodakis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4367-4375

The human visual system has the remarkably ability to be able to effortlessly learn novel concepts from only a few examples. Mimicking the same behavior on machine learning vision systems is an interesting and very challenging research problem with many practical advantages on real world vision applications. In this context, the goal of our work is to devise a few-shot visual learning system that during test time it will be able to efficiently learn novel categories from only a few training data while at the same time it will not forget the initial categories on which it was trained (here called base categories). To achieve that goal we propose (a) to extend an object recognition system with an attention based few-shot classification weight generator, and (b) to redesign the classifier of a ConvNet model as the cosine similarity function between feature representations and classification weight vectors. The latter, apart from unifying the recognition of both novel and base categories, it also leads to feature representations that generalize better on "unseen" categories. We extensively evaluate our approach on Mini-ImageNet where we manage to improve the prior state-of-the-art on few-shot recognition (i.e., we achieve 56.20% and 73.00% on the 1-shot and 5-shot settings respectively) while at the same time we do not sacrifice any accuracy on the base categories, which is a characteristic that most prior approaches lack. Finally, we apply our approach on the recently introduced few-shot benchmark of Bharath and Girshick where we also achieve state-of-the-art results.
****

Two-Step Quantization for Low-Bit Neural Networks

Peisong Wang, Qinghao Hu, Yifan Zhang, Chunjie Zhang, Yang Liu, Jian Cheng; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4376-4384

Every bit matters in the hardware design of quantized neural networks. However, extremely-low-bit representation usually causes large accuracy drop. Thus, how to train extremely-low-bit neural networks with high accuracy is of central importance. Most existing network quantization approaches learn transformations (low-bit weights) as well as encodings (low-bit activations) simultaneously. This tight coupling makes the optimization problem difficult, and thus prevents the network from learning optimal representations. In this paper, we propose a simple yet effective Two-Step Quantization (TSQ) framework, by decomposing the network quantization problem into two steps: code learning and transformation function learning based on the learned codes. For the first step, we propose the sparse quantization method for code learning. The second step can be formulated as a non-li near least square regression problem with low-bit constraints, which can be solved efficiently in an iterative manner. Extensive experiments on CIFAR-10 and ILSVRC-12 datasets demonstrate that the proposed TSQ is effective and outperforms the state-of-the-art by a large margin. Especially, for 2-bit activation and ternary weight quantization of AlexNet, the accuracy of our TSQ drops only about 0.5 points compared with the full-precision counterpart, outperforming current state-of-the-art by more than 5 points.

****

Improved Lossy Image Compression With Priming and Spatially Adaptive Bit Rates for Recurrent Networks

Nick Johnston, Damien Vincent, David Minnen, Michele Covell, Saurabh Singh, Troy Chinen, Sung Jin Hwang, Joel Shor, George Toderici; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4385-4393

We propose a method for lossy image compression based on recurrent, convolutional neural networks that outper- forms BPG (4:2:0), WebP, JPEG2000, and JPEG as mea- sured by MS-SSIM. We introduce three improvements over previous research that lead to this state-of-the-art result us- ing a single model. First, we modify the recurrent architec- ture to improve spatial diffusion, which allows the network to more effectively capture and propagate image informa- tion through the network’s hidden state. Second, in addition to lossless entropy coding, we use a spatially adaptive bit allocation algorithm to more efficiently use the limited num- ber of bits to encode visually complex image regions. Fi- nally, we show that training with a pixel-wise loss weighted by SSIM increases reconstruction quality according to sev- eral metrics. We evaluate our method on the Kodak and Tecnick image sets and compare against standard codecs as well as recently published methods based on deep neural networks.

****

Conditional Probability Models for Deep Image Compression

Fabian Mentzer, Eirikur Agustsson, Michael Tschannen, Radu Timofte, Luc Van Gool; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4394-4402

Deep Neural Networks trained as image auto-encoders have recently emerged as a promising direction for advancing the state-of-the-art in image compression. The key challenge in learning such networks is twofold: To deal with quantization, and to control the trade-off between reconstruction error (distortion) and entropy (rate) of the latent image representation. In this paper, we focus on the latter challenge and propose a new technique to navigate the rate-distortion trade-off for an image compression auto-encoder. The main idea is to directly model the entropy of the latent representation by using a context model: A 3D-CNN which learns a conditional probability model of the latent distribution of the auto-encoder. During training, the auto-encoder makes use of the context model to estimate the entropy of its representation, and the context model is concurrently updated to learn the dependencies between the symbols in the latent representation. Our experiments show that this approach, when measured in MS-SSIM, yields a state-of-the-art image compression system based on a simple convolutional auto-encoder.

****

Deep Diffeomorphic Transformer Networks

Nicki Skafte Detlefsen, Oren Freifeld, Søren Hauberg; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4403-4412

Spatial Transformer layers allow neural networks, at least in principle, to be invariant to large spatial transformations in image data. The model has, however, seen limited uptake as most practical implementations support only transformations that are too restricted, e.g. affine or homographic maps, and/or destructive maps, such as thin plate splines. We investigate the use of flexible diffeomorphic image transformations within such networks and demonstrate that significant performance gains can be attained over currently-used models. The learned transformations are found to be both simple and intuitive, thereby providing insights into individual problem domains. With the proposed framework, a standard convolutional neural network matches state-of-the-art results on face verification with o nly two extra lines of simple TensorFlow code.

****

The Lovász-Softmax Loss: A Tractable Surrogate for the Optimization of the Intersection-Over-Union Measure in Neural Networks

Maxim Berman, Amal Rannen Triki, Matthew B. Blaschko; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4413-4421

The Jaccard index, also referred to as the intersection-over-union score, is commonly employed in the evaluation of image segmentation results given its perceptual qualities, scale invariance - which lends appropriate relevance to small objects, and appropriate counting of false negatives, in comparison to per-pixel losses. We present a method for direct optimization of the mean intersection-over-union loss in neural networks, in the context of semantic image segmentation, based on the convex Lovász extension of submodular losses. The loss is shown to perform better with respect to the Jaccard index measure than the traditionally used cross-entropy loss. We show quantitative and qualitative differences between optimizing the Jaccard index per image versus optimizing the Jaccard index taken over an entire dataset. We evaluate the impact of our method in a semantic segmentation pipeline and show substantially improved intersection-over-union segmentation scores on the Pascal VOC and Cityscapes datasets using state-of-the-art deep learning segmentation architectures.

****

Generative Adversarial Perturbations

Omid Poursaeed, Isay Katsman, Bicheng Gao, Serge Belongie; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4422-4431

In this paper, we propose novel generative models for creating adversarial examples, slightly perturbed images resembling natural images but maliciously crafted to fool pre-trained models. We present trainable deep neural networks for transforming images to adversarial perturbations. Our proposed models can produce image-agnostic and image-dependent perturbations for targeted and non-targeted attacks. We also demonstrate that similar architectures can achieve impressive results in fooling both classification and semantic segmentation models, obviating the need for hand-crafting attack methods for each task. Using extensive experiments on challenging high-resolution datasets such as ImageNet and Cityscapes, we show that our perturbations achieve high fooling rates with small perturbation norms. Moreover, our attacks are considerably faster than current iterative methods at inference time.

****

Learning Strict Identity Mappings in Deep Residual Networks

Xin Yu, Zhiding Yu, Srikumar Ramalingam; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4432-4440

A family of super deep networks, referred to as residual networks or ResNet~cite{he2016deep}, achieved record-beating performance in various visual tasks such as image recognition, object detection, and semantic segmentation. The ability to train very deep networks naturally pushed the researchers to use enormous resources to achieve the best performance. Consequently, in many applications super deep residual networks were employed for just a marginal improvement in performance. In this paper, we propose $epsilon$-ResNet that allows us to automatically discard redundant layers, which produces responses that are smaller than a threshold $epsilon$, without any loss in performance. The $epsilon$-ResNet architecture can be achieved using a few additional rectified linear units in the original ResNet. Our method does not use any additional variables nor numerous trials like other hyper-parameter optimization techniques. The layer selection is achieved using a single training process and the evaluation is performed on CIFAR-10, CIFAR-100, SVHN, and ImageNet datasets. In some instances, we achieve about 80% reduction in the number of parameters.

****

Geometric Robustness of Deep Networks: Analysis and Improvement

Can Kanbak, Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4441

-4449
Deep convolutional neural networks have been shown to be vulnerable to arbitrary geometric transformations. However, there is no systematic method to measure the invariance properties of deep networks to such transformations. We propose ManiFool as a simple yet scalable algorithm to measure the invariance of deep networks. In particular, our algorithm measures the robustness of deep networks to geometric transformations in a worst-case regime as they can be problematic for sensitive applications. Our extensive experimental results show that ManiFool can be used to measure the invariance of fairly complex networks on high dimensional datasets and these values can be used for analyzing the reasons for it. Furthermore, we build on ManiFool to propose a new adversarial training scheme and we show its effectiveness on improving the invariance properties of deep neural networks.
****

View Extrapolation of Human Body From a Single Image
Hao Zhu, Hao Su, Peng Wang, Xun Cao, Ruigang Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4450-4459
We study how to synthesize novel views of human body from a single image. Though recent deep learning based methods work well for rigid objects, they often fail on objects with large articulation, like human bodies. The core step of existing methods is to fit a map from the observable views to novel views by CNNs; however, the rich articulation modes of human body make it rather challenging for CNNs to memorize and interpolate the data well. To address the problem, we propose a novel deep learning based pipeline that explicitly estimates and leverages the geometry of the underlying human body. Our new pipeline is a composition of a shape estimation network and an image generation network, and at the interface a perspective transformation is applied to generate a forward flow for pixel value transportation. Our design is able to factor out the space of data variation and makes learning at each step much easier. Empirically, we show that the performance for pose-varying objects can be improved dramatically. Our method can also be applied on real data captured by 3D sensors, and the flow generated by our methods can be used for generating high quality results in higher resolution.
****

Geometry Aware Constrained Optimization Techniques for Deep Learning
Soumava Kumar Roy, Zakaria Mhammedi, Mehrtash Harandi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4460-4469
In this paper, we generalize the Stochastic Gradient Descent (SGD) and RMSProp algorithms to the setting of Riemannian optimization. SGD is a popular method for large scale optimization. In particular, it is widely used to train the weights of Deep Neural Networks. However, gradients computed using standard SGD can have large variance, which is detrimental for the convergence rate of the algorithm. Other methods such as RMSProp and ADAM address this issue. Nevertheless, these methods cannot be directly applied to constrained optimization problems. In this paper, we extend some popular optimization algorithm to the Riemannian (constrained) setting. We substantiate our proposed extensions with a range of relevant problems in machine learning such as incremental Principal Component Analysis, computating the Riemannian centroids of SPD matrices, and Deep Metric Learning. We achieve competitive results against the state of the art for fine-grained object recognition datasets.
****

PointNetVLAD: Deep Point Cloud Based Retrieval for Large-Scale Place Recognition
Mikaela Angelina Uy, Gim Hee Lee; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4470-4479
Unlike its image based counterpart, point cloud based retrieval for place recognition has remained as an unexplored and unsolved problem. This is largely due to the difficulty in extracting local feature descriptors from a point cloud that can subsequently be encoded into a global descriptor for the retrieval task. In this paper, we propose the PointNetVLAD where we leverage on the recent success of deep networks to solve point cloud based retrieval for place recognition. Specifically, our PointNetVLAD is a combination/modification of the existing Point

Net and NetVLAD, which allows end-to-end training and inference to extract the global descriptor from a given 3D point cloud. Furthermore, we propose the "lazy triplet and quadruplet" loss functions that can achieve more discriminative and generalizable global descriptors to tackle the retrieval task. We create benchmark datasets for point cloud based retrieval for place recognition, and the experimental results on these datasets show the feasibility of our PointNetVLAD.

****

An Efficient and Provable Approach for Mixture Proportion Estimation Using Linear Independence Assumption

Xiyu Yu, Tongliang Liu, Mingming Gong, Kayhan Batmanghelich, Dacheng Tao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4480-4489

In this paper, we study the mixture proportion estimation (MPE) problem in a new setting: given samples from the mixture and the component distributions, we identify the proportions of the components in the mixture distribution. To address this problem, we make use of a linear independence assumption, i.e., the component distributions are independent from each other, which is much weaker than assumptions exploited in the previous MPE methods. Based on this assumption, we propose a method (1) that uniquely identifies the mixture proportions, (2) whose output provably converges to the optimal solution, and (3) that is computationally efficient. We show the superiority of the proposed method over the state-of-the-art methods in two applications including learning with label noise and semi-supervised learning on both synthetic and real-world datasets.

****

VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection

Yin Zhou, Oncel Tuzel; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4490-4499

Accurate detection of objects in 3D point clouds is a central problem in many applications, such as autonomous navigation, housekeeping robots, and augmented/virtual reality. To interface a highly sparse LiDAR point cloud with a region proposal network (RPN), most existing efforts have focused on hand-crafted feature representations, for example, a bird's eye view projection. In this work, we remove the need of manual feature engineering for 3D point clouds and propose VoxelNet, a generic 3D detection network that unifies feature extraction and bounding box prediction into a single stage, end-to-end trainable deep network. Specifically, VoxelNet divides a point cloud into equally spaced 3D voxels and transforms a group of points within each voxel into a unified feature representation through the newly introduced voxel feature encoding (VFE) layer. In this way, the point cloud is encoded as a descriptive volumetric representation, which is then connected to a RPN to generate detections. Experiments on the KITTI car detection benchmark show that VoxelNet outperforms the state-of-the-art LiDAR based 3D detection methods by a large margin. Furthermore, our network learns an effective discriminative representation of objects with various geometries, leading to encouraging results in 3D detection of pedestrians and cyclists, based on only LiDAR.

****

Image to Image Translation for Domain Adaptation

Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, Kyungnam Kim; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4500-4509

We propose a general framework for unsupervised domain adaptation, which allows deep neural networks trained on a source domain to be tested on a different target domain without requiring any training annotations in the target domain. This is achieved by adding extra networks and losses that help regularize the features extracted by the backbone encoder network. To this end we propose the novel use of the recently proposed unpaired image-to-image translation framework to constrain the features extracted by the encoder network. Specifically, we require that the features extracted are able to reconstruct the images in both domains. In addition we require that the distribution of features extracted from images in the two domains are indistinguishable. Many recent works can be seen as specific cases of our general framework. We apply our method for domain adaptation between MNIST, USPS, and SVHN datasets, and Amazon, Webcam and DSLR Office datasets in classification tasks, and also between GTA5 and Cityscapes datasets for a segmentation task. We demonstrate state of the art performance on each of these datasets.

****

MobileNetV2: Inverted Residuals and Linear Bottlenecks

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, Liang-Chieh Chen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4510-4520

In this paper we describe a new mobile architecture, mbox{MobileNetV2}, that improves the state of the art performance of mobile models on multiple tasks and benchmarks as well as across a spectrum of different model sizes. We also describe efficient ways of applying these mobile models to object detection in a novel framework we call mbox{SSDLite}. Additionally, we demonstrate how to build mobile semantic segmentation models through a reduced form of mbox{DeepLabv3} which we call Mobile mbox{DeepLabv3}.   is based on an inverted residual structure where the shortcut connections are between the thin bottleneck layers. The intermediate expansion layer uses lightweight depthwise convolutions to filter features as a source of non-linearity. Additionally, we find that it is important to remove non-linearities in the narrow layers in order to maintain representational power. We demonstrate that this improves performance and provide an intuition that led to this design. Finally, our approach allows decoupling of the input/output domains from the expressiveness of the transformation, which provides a convenient framework for further analysis. We measure our performance on mbox{ImageNet}~cite{Russakovsky:2015:ILS:2846547.2846559} classification, COCO object detection cite{COCO}, VOC image segmentation cite{PASCAL}. We evaluate the trade-offs between accuracy, and number of operations measured by multiply-adds (MAdd), as well as actual latency, and the number of parameters.

****

Im2Struct: Recovering 3D Shape Structure From a Single RGB Image

Chengjie Niu, Jun Li, Kai Xu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4521-4529

We propose to recover 3D shape structures from single RGB images, where structure refers to shape parts represented by cuboids and part relations encompassing connectivity and symmetry. Given a single 2D image with an object depicted, our goal is automatically recover a cuboid structure of the object parts as well as their mutual relations. We develop a convolutional-recursive auto-encoder comprised of structure parsing of a 2D image followed by structure recovering of a cuboid hierarchy. The encoder is achieved by a multi-scale convolutional network trained with the task of shape contour estimation, thereby learning to discern object structures in various forms and scales. The decoder fuses the features of the structure parsing network and the original image, and recursively decodes a hierarchy of cuboids. Since the decoder network is learned to recover part relations including connectivity and symmetry explicitly, the plausibility and generality of part structure recovery can be ensured. The two networks are jointly trained using the training data of contour-mask and cuboid-structure pairs. Such pairs are generated by rendering stock 3D CAD models coming with part segmentation. Our method achieves unprecedentedly faithful and detailed recovery of diverse 3D part structures from single-view 2D images. We demonstrate two applications of our method including structure-guided completion of 3D volumes reconstructed from single-view images and structure-aware interactive editing of 2D images.

****

Trust Your Model: Light Field Depth Estimation With Inline Occlusion Handling

Hendrik Schilling, Maximilian Diebold, Carsten Rother, Bernd Jähne; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4530-4538

We address the problem of depth estimation from light-field images. Our main contribution is a new way to handle occlusions which improves general accuracy and quality of object borders. In contrast to all prior work we work with a model wh ich directly incorporates both depth and occlusion, using a local optimization scheme based on the PatchMatch algorithm. The key benefit of this joint approach is that we utilize all available data, and not erroneously discard valuable information in pre-processing steps. We see the benefit of our approach not only at improved object boundaries, but also at smooth surface reconstruction, where we outperform even methods which focus on good surface regularization. We have evaluated our method on a public light-field dataset, where we achieve state-of-the-art results in nine out of twelve error metrics, with a close tie for the remaining three.

****

Baseline Desensitizing in Translation Averaging

Bingbing Zhuang, Loong-Fah Cheong, Gim Hee Lee; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4539-4547

Many existing translation averaging algorithms are either sensitive to disparate camera baselines and have to rely on extensive preprocessing to improve the observed Epipolar Geometry graph, or if they are robust against disparate camera baselines, require complicated optimization to minimize the highly nonlinear angular error objective. In this paper, we carefully design a simple yet effective bilinear objective function, introducing a variable to perform the requisite normalization. The objective function enjoys the baseline-insensitive property of the angular error and yet is amenable to simple and efficient optimization by block coordinate descent, with good empirical performance. A rotation-assisted Iterative Reweighted Least Squares scheme is further put forth to help deal with outliers. We also contribute towards a better understanding of the behavior of two recent convex algorithms, LUD and Shapefit/kick, clarifying the underlying subtle difference that leads to the performance gap. Finally, we demonstrate that our algorithm achieves overall superior accuracies in benchmark dataset compared to state-of-the-art methods, and is also several times faster.

****

Mining Point Cloud Local Structures by Kernel Correlation and Graph Pooling

Yiru Shen, Chen Feng, Yaoqing Yang, Dong Tian; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4548-4557

Unlike on images, semantic learning on 3D point clouds using a deep network is challenging due to the naturally unordered data structure. Among existing works, PointNet has achieved promising results by directly learning on point sets. However, it does not take full advantage of a point's local neighborhood that contains fine-grained structural information which turns out to be helpful towards better semantic learning. In this regard, we present two new operations to improve PointNet with a more efficient exploitation of local structures. The first one focuses on local 3D geometric structures. In analogy to a convolution kernel for images, we define a point-set kernel as a set of learnable 3D points that jointly respond to a set of neighboring data points according to their geometric affinities measured by kernel correlation, adapted from a similar technique for point cloud registration. The second one exploits local high-dimensional feature structures by recursive feature aggregation on a nearest-neighbor-graph computed from 3D positions. Experiments show that our network can efficiently capture local information and robustly achieve better performances on major datasets. Our code is available at http://www.merl.com/research/license#KCNet

****

Large-Scale Point Cloud Semantic Segmentation With Superpoint Graphs

Loic Landrieu, Martin Simonovsky; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4558-4567

We propose a novel deep learning-based framework to tackle the challenge of semantic segmentation of large-scale point clouds of millions of points. We argue that the organization of 3D point clouds can be efficiently captured by a structure called superpoint graph (SPG), derived from a partition of the scanned scene into geometrically homogeneous elements. SPGs offer a compact yet rich representation of contextual relationships between object parts, which is then exploited by a graph convolutional network. Our framework sets a new state of the art for segmenting outdoor LiDAR scans (+11.9 and +8.8 mIoU points for both Semantic3D te st sets), as well as indoor scans (+12.4 mIoU points for the S3DIS dataset).

****

Very Large-Scale Global SfM by Distributed Motion Averaging

Siyu Zhu, Runze Zhang, Lei Zhou, Tianwei Shen, Tian Fang, Ping Tan, Long Quan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4568-4577

Global Structure-from-Motion (SfM) techniques have demonstrated superior efficiency and accuracy than the conventional incremental approach in many recent studies. This work proposes a divide-and-conquer framework to solve very large global SfM at the scale of millions of images. Specifically, we first divide all images into multiple partitions that preserve strong data association for well posed and parallel local motion averaging. Then, we solve a global motion averaging that determines cameras at partition boundaries and a similarity transformation per partition to register all cameras in a single coordinate frame. Finally, local and global motion averaging are iterated until convergence. Since local camera poses are fixed during the global motion average, we can avoid caching the whole reconstruction in memory at once. This distributed framework significantly enhances the efficiency and robustness of large-scale motion averaging.

****

ScanComplete: Large-Scale Scene Completion and Semantic Segmentation for 3D Scans

Angela Dai, Daniel Ritchie, Martin Bokeloh, Scott Reed, Jürgen Sturm, Matthias Nießner; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4578-4587

We introduce ScanComplete, a novel data-driven approach for taking an incomplete 3D scan of a scene as input and predicting a complete 3D model along with per-voxel semantic labels. The key contribution of our method is its ability to handle large scenes with varying spatial extent, managing the cubic growth in data size as scene size increases. To this end, we devise a fully-convolutional generative 3D CNN model whose filter kernels are invariant to the overall scene size. The model can be trained on scene subvolumes but deployed on arbitrarily large scenes at test time. In addition, we propose a coarse-to-fine inference strategy in order to produce high-resolution output while also leveraging large input context sizes. In an extensive series of experiments, we carefully evaluate different model design choices, considering both deterministic and probabilistic models for completion and semantic inference. Our results show that we outperform other methods not only in the size of the environments handled and processing efficiency, but also with regard to completion quality and semantic segmentation performance by a significant margin.

****

Solving the Perspective-2-Point Problem for Flying-Camera Photo Composition

Ziquan Lan, David Hsu, Gim Hee Lee; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4588-4596

Drone-mounted flying cameras will revolutionize photo-taking. The user, instead of holding a camera in hand and manually searching for a viewpoint, will interact directly with image contents in the viewfinder through simple gestures, and the flying camera will achieve the desired viewpoint through the autonomous flying capability of the drone. This work studies the underlying viewpoint search problem for composing a photo with two objects of interest, a common situation in photo-taking. We model it as a Perspective-2-Point (P2P) problem, which is under-constrained to determine the six degrees-of-freedom camera pose uniquely. By incorporating the user's composition requirements and minimizing the camera's flying distance, we form a constrained nonlinear optimization problem and solve it in closed form. Experiments on synthetic data sets and on a real flying camera system indicate promising results.

****

Reflection Removal for Large-Scale 3D Point Clouds

Jae-Seong Yun, Jae-Young Sim; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4597-4605

Large-scale 3D point clouds (LS3DPCs) captured by terrestrial LiDAR scanners oft en exhibit reflection artifacts by glasses, which degrade the performance of related computer vision techniques. In this paper, we propose an efficient reflection removal algorithm for LS3DPCs. We first partition the unit sphere into local surface patches which are then classified into the ordinary patches and the glass patches according to the number of echo pulses from emitted laser pulses. Then we estimate the glass region of dominant reflection artifacts by measuring the reliability. We also detect and remove the virtual points using the conditions of the reflection symmetry and the geometric similarity. We test the performance of the proposed algorithm on LS3DPCs capturing real-world outdoor scenes, and show that the proposed algorithm estimates valid glass regions faithfully and removes the virtual points caused by reflection artifacts successfully.

****

Attentional ShapeContextNet for Point Cloud Recognition

Saining Xie, Sainan Liu, Zeyu Chen, Zhuowen Tu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4606-4615

We tackle the problem of point cloud recognition. Unlike previous approaches where a point cloud is either converted into a volume/image or represented independently in a permutation-invariant set, we develop a new representation by adopting the concept of shape context as the building block in our network design. The resulting model, called ShapeContextNet, consists of a hierarchy with modules not relying on a fixed grid while still enjoying properties similar to those in convolutional neural networks --- being able to capture and propagate the object part information. In addition, we find inspiration from self-attention based models to include a simple yet effective contextual modeling mechanism --- making the contextual region selection, the feature aggregation, and the feature transformation process fully automatic. ShapeContextNet is an end-to-end model that can be applied to the general point cloud classification and segmentation problems. We observe competitive results on a number of benchmark datasets.

****

Geometry-Aware Deep Network for Single-Image Novel View Synthesis

Miaomiao Liu, Xuming He, Mathieu Salzmann; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4616-4624

This paper tackles the problem of novel view synthesis from a single image. In particular, we target real-world scenes with rich geometric structure, a challenging task due to the large appearance variations of such scenes and the lack of simple 3D models to represent them. Modern, learning-based approaches mostly focus on appearance to synthesize novel views and thus tend to generate predictions that are inconsistent with the underlying scene structure. By contrast, in this paper, we propose to exploit the 3D geometry of the scene to synthesize a novel view. Specifically, we approximate a real-world scene by a fixed number of planes, and learn to predict a set of homographies and their corresponding region masks to transform the input image into a novel view. To this end, we develop a new region-aware geometric transform network that performs these multiple tasks in a common framework. Our results on the outdoor KITTI and the indoor ScanNet datasets demonstrate the effectiveness of our network to generate high-quality synthetic views that respect the scene geometry, thus outperforming the state-of-the-art methods.

****

InverseFaceNet: Deep Monocular Inverse Face Rendering

Hyeongwoo Kim, Michael Zollhöfer, Ayush Tewari, Justus Thies, Christian Richardt, Christian Theobalt; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4625-4634

We introduce InverseFaceNet, a deep convolutional inverse rendering framework for faces that jointly estimates facial pose, shape, expression, reflectance and illumination from a single input image. By estimating all parameters from just a single image, advanced editing possibilities on a single face image, such as appearance editing and relighting, become feasible in real time. Most previous learning-based face reconstruction approaches do not jointly recover all dimensions, or are severely limited in terms of visual quality. In contrast, we propose to recover high-quality facial pose, shape, expression, reflectance and illuminatio n using a deep neural network that is trained using a large, synthetically created training corpus. Our approach builds on a novel loss function that measures model-space similarity directly in parameter space and significantly improves reconstruction accuracy.We further propose a self-supervised bootstrapping process in the network training loop, which iteratively updates the synthetic training corpus to better reflect the distribution of real-world imagery. We demonstrate that this strategy outperforms completely synthetically trained networks. Finally, we show high-quality reconstructions and compare our approach to several state-of-the-art approaches.

****

Sparse Photometric 3D Face Reconstruction Guided by Morphable Models

Xuan Cao, Zhang Chen, Anpei Chen, Xin Chen, Shiying Li, Jingyi Yu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4635-4644

We present a novel 3D face reconstruction technique that leverages sparse photometric stereo (PS) and latest advances on face registration / modeling from a single image. We observe that 3D morphable faces approach provides a reasonable geometry proxy for light position calibration. Specifically, we develop a robust optimization technique that can calibrate per-pixel lighting direction and illumination at a very high precision without assuming uniform surface albedos. Next, we apply semantic segmentation on input images and the geometry proxy to refine hairy vs. bare skin regions using tailored filter. Experiments on synthetic and real data show that by using a very small set of images, our technique is able to reconstruct fine geometric details such as wrinkles, eyebrows, whelks, pores, etc, comparable to and sometimes surpassing movie quality productions.

****

Texture Mapping for 3D Reconstruction With RGB-D Sensor

Yanping Fu, Qingan Yan, Long Yang, Jie Liao, Chunxia Xiao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4645-4653

Acquiring realistic texture details for 3D models is important in 3D reconstruction. However, the existence of geometric errors, caused by noisy RGB-D sensor data, always makes the color images cannot be accurately aligned onto reconstructed 3D models. In this paper, we propose a global-to-local correction strategy to obtain more desired texture mapping results. Our algorithm first adaptively selects an optimal image for each face of the 3D model, which can effectively remove blurring and ghost artifacts produced by multiple image blending. We then adopt a non-rigid global-to-local correction step to reduce the seaming effect between textures. This can effectively compensate for the texture and the geometric misalignment caused by camera pose drift and geometric errors. We evaluate the proposed algorithm in a range of complex scenes and demonstrate its effective performance in generating seamless high fidelity textures for 3D models.

****

Learning Less Is More - 6D Camera Localization via 3D Surface Regression

Eric Brachmann, Carsten Rother; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4654-4662

Popular research areas like autonomous driving and augmented reality have renewed the interest in image-based camera localization. In this work, we address the task of predicting the 6D camera pose from a single RGB image in a given 3D environment. With the advent of neural networks, previous works have either learned the entire camera localization process, or multiple components of a camera localization pipeline. Our key contribution is to demonstrate and explain that learning a single component of this pipeline is sufficient. This component is a fully convolutional neural network for densely regressing so-called scene coordinates, defining the correspondence between the input image and the 3D scene space. The neural network is prepended to a new end-to-end trainable pipeline. Our system is efficient, highly accurate, robust in training, and exhibits outstanding generalization capabilities. It exceeds state-of-the-art consistently on indoor and outdoor datasets. Interestingly, our approach surpasses existing techniques even without utilizing a 3D model of the scene during training, since the network is

able to discover 3D scene geometry automatically, solely from single-view constraints.
****

Feature Mapping for Learning Fast and Accurate 3D Pose Inference From Synthetic Images

Mahdi Rad, Markus Oberweger, Vincent Lepetit; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4663-4672

We propose a simple and efficient method for exploiting synthetic images when training a Deep Network to predict a 3D pose from an image. The ability of using synthetic images for training a Deep Network is extremely valuable as it is easy to create a virtually infinite training set made of such images, while capturing and annotating real images can be very cumbersome. However, synthetic images do not resemble real images exactly, and using them for training can result in suboptimal performance. It was recently shown that for exemplar-based approaches, it is possible to learn a mapping from the exemplar representations of real images to the exemplar representations of synthetic images. In this paper, we show that this approach is more general, and that a network can also be applied after the mapping to infer a 3D pose: At run-time, given a real image of the target object, we first compute the features for the image, map them to the feature space of synthetic images, and finally use the resulting features as input to another network which predicts the 3D pose. Since this network can be trained very effectively by using synthetic images, it performs very well in practice, and inference is faster and more accurate than with an exemplar-based approach. We demonstrate our approach on the LINEMOD dataset for 3D object pose estimation from color images, and the NYU dataset for 3D hand pose estimation from depth maps. We show that it allows us to outperform the state-of-the-art on both datasets.
****

Indoor RGB-D Compass From a Single Line and Plane

Pyojin Kim, Brian Coltin, H. Jin Kim; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4673-4680

We propose a novel approach to estimate the three degrees of freedom (DoF) drift-free rotational motion of an RGB-D camera from only a single line and plane in the Manhattan world (MW). Previous approaches exploit the surface normal vectors and vanishing points to achieve accurate 3-DoF rotation estimation. However, they require multiple orthogonal planes or many consistent lines to be visible throughout the entire rotation estimation process; otherwise, these approaches fail. To overcome these limitations, we present a new method that estimates absolute camera orientation from only a single line and a single plane in RANSAC, which corresponds to the theoretical minimal sampling for 3-DoF rotation estimation. Once we find an initial rotation estimate, we refine the camera orientation by minimizing the average orthogonal distance from the endpoints of the lines parallel to the MW axes. We demonstrate the effectiveness of the proposed algorithm through an extensive evaluation on a variety of RGB-D datasets and compare with other state-of-the-art methods.
****

Geometry-Aware Network for Non-Rigid Shape Prediction From a Single View

Albert Pumarola, Antonio Agudo, Lorenzo Porzi, Alberto Sanfeliu, Vincent Lepetit, Francesc Moreno-Noguer; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4681-4690

We propose a method for predicting the 3D shape of a deformable surface from a single view. By contrast with previous approaches, we do not need a pre-registered template of the surface, and our method is robust to the lack of texture and partial occlusions. At the core of our approach is a geometry-aware deep architecture that tackles the problem as usually done in analytic solutions: first perform 2D detection of the mesh and then estimate a 3D shape that is geometrically consistent with the image. We train this architecture in an end-to-end manner using a large dataset of synthetic renderings of shapes under different levels of deformation, material properties, textures and lighting conditions. We evaluate our approach on a test split of this dataset and available real benchmarks, consistently improving state-of-the-art solutions with a significantly lower computat ional time.

****

Sim2Real Viewpoint Invariant Visual Servoing by Recurrent Control

Fereshteh Sadeghi, Alexander Toshev, Eric Jang, Sergey Levine; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4691-4699

Humans are remarkably proficient at controlling their limbs and tools from a wide range of viewpoints. In robotics, this ability is referred to as visual servoing: moving a tool or end-point to a desired location using primarily visual feedback. In this paper, we propose learning viewpoint invariant visual servoing skills in a robot manipulation task. We train a deep recurrent controller that can automatically determine which actions move the end-effector of a robotic arm to a desired object. This problem is fundamentally ambiguous: under severe variation in viewpoint, it may be impossible to determine the actions in a single feedforward operation. Instead, our visual servoing approach uses its memory of past movements to understand how the actions affect the robot motion from the current viewpoint, correcting mistakes and gradually moving closer to the target. This ability is in stark contrast to previous visual servoing methods, which assume known dynamics or require a calibration phase. We learn our recurrent controller using simulated data, synthetic demonstrations and reinforcement learning. We then describe how the resulting model can be transferred to a real-world robot by disentangling perception from control and only adapting the visual layers. The adapted model can servo to previously unseen objects from novel viewpoints on a real-world Kuka IIWA robotic arm. For supplementary videos, see: href{https://www.youtube.com/watch?v=oLgM2Bnb7fo}{https://www.youtube.com/watch?v=oLgM2Bnb7fo}

****

DocUNet: Document Image Unwarping via a Stacked U-Net

Ke Ma, Zhixin Shu, Xue Bai, Jue Wang, Dimitris Samaras; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4700-4709

Capturing document images is a common way for digitizing and recording physical documents due to the ubiquitousness of mobile cameras. To make text recognition easier, it is often desirable to digitally flatten a document image when the physical document sheet is folded or curved. In this paper, we develop the first learning-based method to achieve this goal. We propose a stacked U-Net with intermediate supervision to directly predict the forward mapping from a distorted image to its rectified version. Because large-scale real-world data with ground truth deformation is difficult to obtain, we create a synthetic dataset with approximately 100 thousand images by warping non-distorted document images. The network is trained on this dataset with various data augmentations to improve its generalization ability. We further create a comprehensive benchmark that covers various real-world conditions. We evaluate the proposed model quantitatively and qualitatively on the proposed benchmark, and compare it with previous non-learning-based methods.

****

Analysis of Hand Segmentation in the Wild

Aisha Urooj, Ali Borji; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4710-4719

A large number of works in egocentric vision have concentrated on action and object recognition. Detection and segmentation of hands in first-person videos, however, has less been explored. For many applications in this domain, it is necessary to accurately segment not only hands of the camera wearer but also the hands of others with whom he is interacting. Here, we take an in-depth look at the hand segmentation problem. In the quest for robust hand segmentation methods, we evaluated the performance of the state of the art semantic segmentation methods, off the shelf and fine-tuned, on existing datasets. We fine-tune RefineNet, a leading semantic segmentation method, for hand segmentation and find that it does much better than the best contenders. Existing hand segmentation datasets are collected in the laboratory settings. To overcome this limitation, we contribute by collecting two new datasets: a) EgoYouTubeHands including egocentric videos co ntaining hands in the wild, and b) HandOverFace to analyze the performance of our models in presence of similar appearance occlusions. We further explore whether conditional random fields can help refine generated hand segmentations. To demonstrate the benefit of accurate hand maps, we train a CNN for hand-based activity recognition and achieve higher accuracy when a CNN was trained using hand maps produced by the fine-tuned RefineNet. Finally, we annotate a subset of the EgoHands dataset for fine-grained action recognition and show that an accuracy of 58.6% can be achieved by just looking at a single hand pose which is much better than the chance level (12.5%).

****

RoadTracer: Automatic Extraction of Road Networks From Aerial Images

Favyen Bastani, Songtao He, Sofiane Abbar, Mohammad Alizadeh, Hari Balakrishnan, Sanjay Chawla, Sam Madden, David DeWitt; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4720-4728

Mapping road networks is currently both expensive and labor-intensive. High-resolution aerial imagery provides a promising avenue to automatically infer a road network. Prior work uses convolutional neural networks (CNNs) to detect which pixels belong to a road (segmentation), and then uses complex post-processing heuristics to infer graph connectivity. We show that these segmentation methods have high error rates because noisy CNN outputs are difficult to correct. We propose RoadTracer, a new method to automatically construct accurate road network maps from aerial images. RoadTracer uses an iterative search process guided by a CNN-based decision function to derive the road network graph directly from the output of the CNN. We compare our approach with a segmentation method on fifteen cities, and find that at a 5% error rate, RoadTracer correctly captures 45% more junctions across these cities.

****

Alternating-Stereo VINS: Observability Analysis and Performance Evaluation

Mrinal K. Paul, Stergios I. Roumeliotis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4729-4737

One approach to improve the accuracy and robustness of vision-aided inertial navigation systems (VINS) that employ low-cost inertial sensors, is to obtain scale information from stereoscopic vision. Processing images from two cameras, however, is computationally expensive and increases latency. To address this limitation, in this work, a novel two-camera alternating-stereo VINS is presented. Specifically, the proposed system triggers the left-right cameras in an alternating fashion, estimates the poses corresponding to the left camera only, and introduces a linear interpolation model for processing the alternating right camera measurements. Although not a regular stereo system, the alternating visual observations when employing the proposed interpolation scheme, still provide scale information, as shown by analyzing the observability properties of the vision-only corresponding system. Finally, the performance gain, of the proposed algorithm over its monocular and stereo counterparts is assessed using various datasets.

****

Soccer on Your Tabletop

Konstantinos Rematas, Ira Kemelmacher-Shlizerman, Brian Curless, Steve Seitz; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4738-4747

We present a system that transforms a monocular video of a soccer game into a moving 3D reconstruction, in which the players and field can be rendered interactively with a 3D viewer or through an Augmented Reality device. At the heart of our paper is an approach to estimate the depth map of each player, using a CNN that is trained on 3D player data extracted from soccer video games. We compare with state of the art body pose and depth estimation techniques, and show results on both synthetic ground truth benchmarks, and real YouTube soccer footage.

****

EPINET: A Fully-Convolutional Neural Network Using Epipolar Geometry for Depth From Light Field Images

Changha Shin, Hae-Gon Jeon, Youngjin Yoon, In So Kweon, Seon Joo Kim; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018

, pp. 4748-4757

Light field cameras capture both the spatial and the angular properties of light rays in space. Due to its property, one can compute the depth from light fields in uncontrolled lighting environments, which is a big advantage over active sensing devices. Depth computed from light fields can be used for many applications including 3D modelling and refocusing. However, light field images from hand-held cameras have very narrow baselines with noise, making the depth estimation difficult. Many approaches have been proposed to overcome these limitations for the light field depth estimation, but there is a clear trade-off between the accuracy and the speed in these methods. In this paper, we introduce a fast and accurate light field depth estimation method based on a fully-convolutional neural network. Our network is designed by considering the light field geometry and we also overcome the lack of training data by proposing light field specific data augmentation methods. We achieved the top rank in the HCI 4D Light Field Benchmark on most metrics, and we also demonstrate the effectiveness of the proposed method on real-world light-field images.

****

A Hybrid l1-l0 Layer Decomposition Model for Tone Mapping

Zhetong Liang, Jun Xu, David Zhang, Zisheng Cao, Lei Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4758-4766

Tone mapping aims to reproduce a standard dynamic range image from a high dynamic range image with visual information preserved. State-of-the-art tone mapping algorithms mostly decompose an image into a base layer and a detail layer, and process them accordingly. These methods may have problems of halo artifacts and over-enhancement, due to the lack of proper priors imposed on the two layers. In this paper, we propose a hybrid L1-L0 decomposition model to address these problems. Specifically, an L1 sparsity term is imposed on the base layer to model its piecewise smoothness property. An L0 sparsity term is imposed on the detail layer as a structural prior, which leads to piecewise constant effect. We further propose a multiscale tone mapping scheme based on our layer decomposition model. Experiments show that our tone mapping algorithm achieves visually compelling results with little halo artifacts, outperforming the state-of-the-art tone mapping algorithms in both subjective and objective evaluations.

****

Deeply Learned Filter Response Functions for Hyperspectral Reconstruction

Shijie Nie, Lin Gu, Yinqiang Zheng, Antony Lam, Nobutaka Ono, Imari Sato; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4767-4776

Hyperspectral reconstruction from RGB imaging has recently achieved significant progress via sparse coding and deep learning. However, a largely ignored fact is that existing RGB cameras are tuned to mimic human richromatic perception, thus their spectral responses are not necessarily optimal for hyperspectral reconstruction. In this paper, rather than use RGB spectral responses, we simultaneously learn optimized camera spectral response functions (to be implemented in hardware) and a mapping for spectral reconstruction by using an end-to-end network. Our core idea is that since camera spectral filters act in effect like the convolution layer, their response functions could be optimized by training standard neural networks. We propose two types of designed filters: a three-chip setup without spatial mosaicing and a single-chip setup with a Bayer-style 2x2 filter array. Numerical simulations verify the advantages of deeply learned spectral responses compared to existing RGB cameras. More interestingly, by considering physical restrictions in the design process, we are able to realize the deeply learned spectral response functions by using modern film filter production technologies, and thus construct data-inspired multispectral cameras for snapshot hyperspectral imaging.

****

CRRN: Multi-Scale Guided Concurrent Reflection Removal Network

Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, Alex C. Kot; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4

777-4785

Removing the undesired reflections from images taken through the glass is of broad application to various computer vision tasks. Non-learning based methods utilize different handcrafted priors such as the separable sparse gradients caused by different levels of blurs, which often fail due to their limited description capability to the properties of real-world reflections. In this paper, we propose the Concurrent Reflection Removal Network (CRRN) to tackle this problem in a unified framework. Our network integrates image appearance information and multi-scale gradient information with human perception inspired loss function, and is trained on a new dataset with 3250 reflection images taken under diverse real-world scenes. Extensive experiments on a public benchmark dataset show that the proposed method performs favorably against state-of-the-art methods.

****

Single Image Reflection Separation With Perceptual Losses

Xuaner Zhang, Ren Ng, Qifeng Chen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4786-4794

We present an approach to separating reflection from a single image. The approach uses a fully convolutional network trained end-to-end with losses that exploit low-level and high-level image information. Our loss function includes two perceptual losses: a feature loss from a visual perception network, and an adversarial loss that encodes characteristics of images in the transmission layers. We also propose a novel exclusion loss that enforces pixel-level layer separation. We create a dataset of real-world images with reflection and corresponding ground-truth transmission layers for quantitative evaluation and model training. We validate our method through comprehensive quantitative experiments and show that our approach outperforms state-of-the-art reflection removal methods in PSNR, SSIM, and perceptual user study. We also extend our method to two other image enhancement tasks to demonstrate the generality of our approach.

****

A Robust Method for Strong Rolling Shutter Effects Correction Using Lines With Automatic Feature Selection

Yizhen Lao, Omar Ait-Aider; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4795-4803

We present a robust method which compensates RS distortions in a single image using a set of image curves, basing on the knowledge that they correspond to 3D straight lines. Unlike in existing work, no a priori knowledge about the line directions (e.g. Manhattan World assumption) is required. We first formulate a parametric equation for the projection of a 3D straight line viewed by a moving rolling shutter camera under a uniform motion model. Then we propose a method which efficiently estimates ego angular velocity separately from pose parameters, using at least 4 image curves. Moreover, we propose for the first time a RANSAC-like strategy to select image curves which really correspond to 3D straight lines and reject those corresponding to actual curves in 3D world. A comparative experimental study with both synthetic and real data from famous benchmarks shows that the proposed method outperforms all the existing techniques from the state-of-the-art.

****

Time-Resolved Light Transport Decomposition for Thermal Photometric Stereo

Kenichiro Tanaka, Nobuhiro Ikeya, Tsuyoshi Takatani, Hiroyuki Kubo, Takuya Funatomi, Yasuhiro Mukaigawa; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4804-4813

We present a novel time-resolved light transport decomposition method using thermal imaging. Because the speed of heat propagation is much slower than the speed of light propagation, transient transport of far infrared light can be observed at a video frame rate. A key observation is that the thermal image looks similar to the visible light image in an appropriately controlled environment. This implies that conventional computer vision techniques can be straightforwardly applied to the thermal image. We show that the diffuse component in the thermal image can be separated and, therefore, the surface normals of objects can be estimated by the Lambertian photometric stereo. The effectiveness of our method is eval

uated by conducting real-world experiments, and its applicability to black body, transparent, and translucent objects is shown.
****

Efficient Diverse Ensemble for Discriminative Co-Tracking

Kourosh Meshgi, Shigeyuki Oba, Shin Ishii; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4814-4823

Ensemble discriminative tracking utilizes a committee of classifiers, to label data samples, which are in turn, used for retraining the tracker to localize the target using the collective knowledge of the committee. Committee members could vary in their features, memory update schemes, or training data, however, it is inevitable to have committee members that excessively agree because of large overlaps in their version space. To remove this redundancy and have an effective ensemble learning, it is critical for the committee to include consistent hypotheses that differ from one-another, covering the version space with minimum overlaps. In this study, we propose an online ensemble tracker that directly generates a diverse committee by generating an efficient set of artificial training. The artificial data is sampled from the empirical distribution of the samples taken from both target and background, whereas the process is governed by query-by-committee to shrink the overlap between classifiers. The experimental results demonstrate that the proposed scheme outperforms conventional ensemble trackers on public benchmarks.
****

Rolling Shutter and Radial Distortion Are Features for High Frame Rate Multi-Camera Tracking

Akash Bapat, True Price, Jan-Michael Frahm; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4824-4833

Traditionally, camera-based tracking approaches have treated rolling shutter and radial distortion as imaging artifacts that have to be overcome and corrected for in order to apply standard camera models and scene reconstruction methods. In this paper, we introduce a novel multi-camera tracking approach that for the first time jointly leverages the information introduced by rolling shutter and radial distortion as a feature to achieve superior performance with respect to high-frequency camera pose estimation. In particular, our system is capable of attaining high tracking rates that were previously unachievable. Our approach explicitly leverages rolling shutter capture and radial distortion to process individual rows, rather than entire image frames, for accurate camera motion estimation. We estimate a per-row 6 DoF pose of a rolling shutter camera by tracking multiple points on a radially distorted row whose rays span a curved surface in 3D space. Although tracking systems for rolling shutter cameras exist, we are the first to leverage radial distortion to measure a per-row pose -- enabling us to use less than half the number of cameras required by the previous state of the art. We validate our system on both synthetic and real imagery.
****

A Twofold Siamese Network for Real-Time Object Tracking

Anfeng He, Chong Luo, Xinmei Tian, Wenjun Zeng; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4834-4843

Observing that Semantic features learned in an image classification task and Appearance features learned in a similarity matching task complement each other, we build a twofold Siamese network, named SA-Siam, for real-time object tracking. SA-Siam is composed of a semantic branch and an appearance branch. Each branch is a similarity learning Siamese network. An important design choice in SA-Siam is to separately train the two branches to keep the heterogeneity of the two types of features. In addition, we propose a channel attention mechanism for the semantic branch. Channel-wise weights are computed according to the channel activations around the target position. While the inherited architecture from SiamFC allows our tracker to operate beyond real-time, the twofold design and the attention mechanism significantly improve the tracking performance. The proposed SA-Siam outperforms all other real-time trackers by a large margin on OTB-2013/50/100 benchmarks.
****

Multi-Cue Correlation Filters for Robust Visual Tracking

Ning Wang, Wengang Zhou, Qi Tian, Richang Hong, Meng Wang, Houqiang Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4844-4853

In recent years, many tracking algorithms achieve impressive performance via fusing multiple types of features, however, most of them fail to fully explore the context among the adopted multiple features and the strength of them. In this paper, we propose an efficient multi-cue analysis framework for robust visual tracking. By combining different types of features, our approach constructs multiple experts through Discriminative Correlation Filter (DCF) and each of them tracks the target independently. With the proposed robustness evaluation strategy, the suitable expert is selected for tracking in each frame. Furthermore, the divergence of multiple experts reveals the reliability of the current tracking, which is quantified to update the experts adaptively to keep them from corruption. Through the proposed multi-cue analysis, our tracker with standard DCF and deep features achieves outstanding results on several challenging benchmarks: OTB-2013, OTB-2015, Temple-Color and VOT 2016. On the other hand, when evaluated with only simple hand-crafted features, our method demonstrates comparable performance amongst complex non-realtime trackers, but exhibits much better efficiency, with a speed of 45 FPS on a CPU.

****

Learning Attentions: Residual Attentional Siamese Network for High Performance Online Visual Tracking

Qiang Wang, Zhu Teng, Junliang Xing, Jin Gao, Weiming Hu, Stephen Maybank; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4854-4863

Offline training for object tracking has recently shown great potentials in balancing tracking accuracy and speed. However, it is still difficult to adapt an offline trained model to a target tracked online. This work presents a Residual Attentional Siamese Network (RASNet) for high performance object tracking. The RASNet model reformulates the correlation filter within a Siamese tracking framework, and introduces different kinds of the attention mechanisms to adapt the model without updating the model online. In particular, by exploiting the offline trained general attention, the target adapted residual attention, and the channel favored feature attention, the RASNet not only mitigates the over-fitting problem in deep network training, but also enhances its discriminative capacity and adaptability due to the separation of representation learning and discriminator learning. The proposed deep architecture is trained from end to end and takes full advantage of the rich spatial temporal information to achieve robust visual tracking. Experimental results on two latest benchmarks, OTB-2015 and VOT2017, show that the RASNet tracker has the state-of-the-art tracking accuracy while runs at more than 80 frames per second.

****

SINT++: Robust Visual Tracking via Adversarial Positive Instance Generation

Xiao Wang, Chenglong Li, Bin Luo, Jin Tang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4864-4873

Existing visual trackers are easily disturbed by occlusion,blurandlargedeformation. Inthechallengesofocclusion, motion blur and large object deformation, the performance of existing visual trackers may be limited due to the followingissues: i)Adoptingthedensesamplingstrategyto generate positive examples will make them less diverse; ii) Thetrainingdatawithdifferentchallengingfactorsarelimited, even though through collecting large training dataset. Collecting even larger training dataset is the most intuitive paradigm, but it may still can not cover all situations and the positive samples are still monotonous. In this paper, we propose to generate hard positive samples via adversarial learning for visual tracking. Specifically speaking, we assume the target objects all lie on a manifold, hence, we introduce the positive samples generation network (PSGN) to sampling massive diverse training data through traversing over the constructed target object manifold. The generated diverse target object images can enrich the training dataset and enhance the robustness of visual trackers. To make the tracker more robu st to occlusion, we adopt the hard positive transformation network (HPTN) which can generate hard samples for tracking algorithm to recognize. We train this network with deep reinforcement learning to automaticallyoccludethetargetobjectwith anegativepatch. Based on the generated hard positive samples, we train a Siamese network for visual tracking and our experiments validate the effectiveness of the introduced algorithm.

****

High-Speed Tracking With Multi-Kernel Correlation Filters
Ming Tang, Bin Yu, Fan Zhang, Jinqiao Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4874-4883

Correlation filter (CF) based trackers are currently ranked top in terms of their performances. Nevertheless, only some of them, such as KCF [henriques12&15] and MKCF[tang&Feng15}, are able to exploit the powerful discriminability of non-linear kernels. Although MKCF achieves more powerful discriminability than KCF through introducing multi-kernel learning (MKL) into KCF, its improvement over KCF is quite limited and its computational burden increases significantly in comparison with KCF. In this paper, we will introduce the MKL into KCF in a different way than MKCF. We reformulate the MKL version of CF objective function with its upper bound, alleviating the negative mutual interference of different kernels significantly. Our novel MKCF tracker, MKCFup, outperforms KCF and MKCF with large margins and can still work at very high fps. Extensive experiments on public data sets show that our method is superior to state-of-the-art algorithms for target objects of small move at very high speed.

****

Occlusion Aware Unsupervised Learning of Optical Flow
Yang Wang, Yi Yang, Zhenheng Yang, Liang Zhao, Peng Wang, Wei Xu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4884-4893

It has been recently shown that a convolutional neural network can learn optical flow estimation with unsuper- vised learning. However, the performance of the unsuper- vised methods still has a relatively large gap compared to its supervised counterpart. Occlusion and large motion are some of the major factors that limit the current unsuper- vised learning of optical flow methods. In this work we introduce a new method which models occlusion explicitly and a new warping way that facilitates the learning of large motion. Our method shows promising results on Flying Chairs, MPI-Sintel and KITTI benchmark datasets. Espe- cially on KITTI dataset where abundant unlabeled samples exist, our unsupervised method outperforms its counterpart trained with supervised learning.

****

Revisiting Video Saliency: A Large-Scale Benchmark and a New Model
Wenguan Wang, Jianbing Shen, Fang Guo, Ming-Ming Cheng, Ali Borji; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4894-4903

In this work, we contribute to video saliency research in two ways. First, we introduce a new benchmark for predicting human eye movements during dynamic scene free-viewing, which is long-time urged in this field. Our dataset, named DHF1K~(Dynamic Human Fixation), consists of 1K high-quality, elaborately selected video sequences spanning a large range of scenes, motions, object types and background complexity. Existing video saliency datasets lack variety and generality of common dynamic scenes and fall short in covering challenging situations in unconstrained environments. In contrast, DHF1K~makes a significant leap in terms of scalability, diversity and difficulty, and is expected to boost video saliency modeling. Second, we propose a novel video saliency model that augments the CNN-LSTM network architecture with an attention mechanism to enable fast, end-to-end saliency learning. The attention mechanism explicitly encodes static saliency information, thus allowing LSTM to focus on learning more flexible temporal saliency representation across successive frames. Such a design fully leverages existing large-scale static fixation datasets, avoids overfitting, and significantly improves training efficiency and testing performance. We thoroughly examine the performance of our model, with respect to state-of-the-art saliency models, on three large-scale datasets (i.e., DHF1K, Hollywood2, UCF sports). Experimental results over more than 1.2K testing videos containing 400K frames demonstrate that our model outperforms other competitors.

****

Learning Spatial-Temporal Regularized Correlation Filters for Visual Tracking
Feng Li, Cheng Tian, Wangmeng Zuo, Lei Zhang, Ming-Hsuan Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4904-4913

Discriminative Correlation Filters (DCF) are efficient in visual tracking but suffer from unwanted boundary effects. Spatially Regularized DCF (SRDCF) has been suggested to resolve this issue by enforcing spatial penalty on DCF coefficients, which, inevitably, improves the tracking performance at the price of increasing complexity. To tackle online updating, SRDCF formulates its model on multiple training images, further adding difficulties in improving efficiency. In this work, by introducing temporal regularization to SRDCF with single sample, we present our spatial-temporal regularized correlation filters (STRCF). The STRCF formulation can not only serve as a reasonable approximation to SRDCF with multiple training samples, but also provide a more robust appearance model than SRDCF in the case of large appearance variations. Besides, it can be efficiently solved via the alternating direction method of multipliers (ADMM). By incorporating both temporal and spatial regularization, our STRCF can handle boundary effects without much loss in efficiency and achieve superior performance over SRDCF in terms of accuracy and speed. Compared with SRDCF, STRCF with hand-crafted features provides a 5× speedup and achieves a gain of 5.4% and 3.6% AUC score on OTB-2015 and Temple-Color, respectively. Moreover, STRCF with deep features also performs favorably against state-of-the-art trackers and achieves an AUC score of 68.3% on OTB-2015.

****

Multimodal Visual Concept Learning With Weakly Supervised Techniques
Giorgos Bouritsas, Petros Koutras, Athanasia Zlatintsi, Petros Maragos; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4914-4923

Despite the availability of a huge amount of video data accompanied by descriptive texts, it is not always easy to exploit the information contained in natural language in order to automatically recognize video concepts. Towards this goal, in this paper we use textual cues as means of supervision, introducing two weakly supervised techniques that extend the Multiple Instance Learning (MIL) framework: the Fuzzy Sets Multiple Instance Learning (FSMIL) and the Probabilistic Labels Multiple Instance Learning (PLMIL). The former encodes the spatio-temporal imprecision of the linguistic descriptions with Fuzzy Sets, while the latter models different interpretations of each description’s semantics with Probabilistic Labels, both formulated through a convex optimization algorithm. In addition, we provide a novel technique to extract weak labels in the presence of complex semantics, that consists of semantic similarity computations. We evaluate our methods on two distinct problems, namely face and action recognition, in the challenging and realistic setting of movies accompanied by their screenplays, contained in the COGNIMUSE database. We show that, on both tasks, our method considerably outperforms a state-of-the-art weakly supervised approach, as well as other baselines.

****

Efficient Large-Scale Approximate Nearest Neighbor Search on OpenCL FPGA
Jialiang Zhang, Soroosh Khoram, Jing Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4924-4932

We present a new method for Product Quantization (PQ) based approximated nearest neighbor search (ANN) in high dimensional spaces. Specifically, we first propose a quantization scheme for the codebook of coarse quantizer, product quantizer, and rotation matrix, to reduce the cost of accessing these codebooks. Our approach also combines a highly parallel k-selection method, which can be fused with the distance calculation to reduce the memory overhead. We implement the proposed method on Intel HARPv2 platform using OpenCL-FPGA. The proposed method signif icantly outperforms state-of-the-art methods on CPU and GPU for high dimensional nearest neighbor queries on billion-scale datasets in terms of query time and accuracy regardless of the batch size. To our best knowledge, this is the first work to demonstrate FPGA performance superior to CPU and GPU on high-dimensional, large-scale ANN datasets.

****

Learning a Complete Image Indexing Pipeline

Himalaya Jain, Joaquin Zepeda, Patrick Pérez, Rémi Gribonval; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4933-4941

To work at scale, a complete image indexing system comprises two components: An inverted file index to restrict the actual search to only a subset that should contain most of the items relevant to the query; An approximate distance computation mechanism to rapidly scan these lists. While supervised deep learning has recently enabled improvements to the latter, the former continues to be based on unsupervised clustering in the literature. In this work, we propose a first system that learns both components within a unifying neural framework of structured binary encoding.

****

Transparency by Design: Closing the Gap Between Performance and Interpretability in Visual Reasoning

David Mascharka, Philip Tran, Ryan Soklaski, Arjun Majumdar; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4942-4950

Visual question answering requires high-order reasoning about an image, which is a fundamental capability needed by machine systems to follow complex directives. Recently, modular networks have been shown to be an effective framework for performing visual reasoning tasks. While modular networks were initially designed with a degree of model transparency, their performance on complex visual reasoning benchmarks was lacking. Current state-of-the-art approaches do not provide an effective mechanism for understanding the reasoning process. In this paper, we close the performance gap between interpretable models and state-of-the-art visual reasoning methods. We propose a set of visual-reasoning primitives which, when composed, manifest as a model capable of performing complex reasoning tasks in an explicitly-interpretable manner. The fidelity and interpretability of the primitives' outputs enable an unparalleled ability to diagnose the strengths and weaknesses of the resulting model. Critically, we show that these primitives are highly performant, achieving state-of-the-art accuracy of 99.1% on the CLEVR dataset. We also show that our model is able to effectively learn generalized representations when provided a small amount of data containing novel object attributes. Using the CoGenT generalization task, we show more than a 20 percentage point improvement over the current state of the art.

****

Fooling Vision and Language Models Despite Localization and Attention Mechanism

Xiaojun Xu, Xinyun Chen, Chang Liu, Anna Rohrbach, Trevor Darrell, Dawn Song; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4951-4961

Adversarial attacks are known to succeed on classifiers, but it has been an open question whether more complex vision systems are vulnerable. In this paper, we study adversarial examples for vision and language models, which incorporate natural language understanding and complex structures such as attention, localization, and modular architectures. In particular, we investigate attacks on a dense captioning model and on two visual question answering (VQA) models. Our evaluation shows that we can generate adversarial examples with a high success rate (i.e., >90%) for these models. Our work sheds new light on understanding adversarial attacks on vision systems which have a language component and shows that attention, bounding box localization, and compositional internal structures are vulnerable to adversarial attacks. These observations will inform future work towards building effective defenses.

****

Categorizing Concepts With Basic Level for Vision-to-Language

Hanzhang Wang, Hanli Wang, Kaisheng Xu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4962-4970

Vision-to-language tasks require a unified semantic understanding of visual content. However, the information contained in image/video is essentially ambiguous on two perspectives manifested on the diverse understanding among different persons and the various understanding grains even for the same person. Inspired by the basic level in early cognition, a Basic Concept (BaC) category is proposed in this work that contains both consensus and proper level of visual content to help neural network tackle the above problems. Specifically, a salient concept category is firstly generated by intersecting the labels of ImageNet and the vocabulary of MSCOCO dataset. Then, according to the observation from human early cognition that children make fewer mistakes on the basic level, the salient category is further refined by clustering concepts with a defined confusion degree which measures the difficulty for convolutional neural network to distinguish class pairs. Finally, a pre-trained model based on GoogLeNet is produced with the proposed BaC category of 1,372 concept classes. To verify the effectiveness of the proposed categorizing method for vision-to-language tasks, two kinds of experiments are performed including image captioning and visual question answering with the benchmark datasets of MSCOCO, Flickr30k and COCO-QA. The experimental results demonstrate that the representations derived from the cognition-inspired BaC category promote representation learning of neural networks on vision-to-language tasks, and a performance improvement is gained without modifying standard models.

****

Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, Aniruddha Kembhavi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4971-4980

A number of studies have found that today's Visual Question Answering (VQA) models are heavily driven by superficial correlations in the training data and lack sufficient image grounding. To encourage development of models geared towards the latter, we propose a new setting for VQA where for every question type, train and test sets have different prior distributions of answers. Specifically, we present new splits of the VQA v1 and VQA v2 datasets, which we call Visual Question Answering under Changing Priors (VQA-CP v1 and VQA-CP v2 respectively). First, we evaluate several existing VQA models under this new setting and show that their performance degrades significantly compared to the original VQA setting. Second, we propose a novel Grounded Visual Question Answering model (GVQA) that contains inductive biases and restrictions in the architecture specifically designed to prevent the model from 'cheating' by primarily relying on priors in the training data. Specifically, GVQA explicitly disentangles the recognition of visual concepts present in the image from the identification of plausible answer space for a given question, enabling the model to more robustly generalize across different distributions of answers. GVQA is built off an existing VQA model -- Stacked Attention Networks (SAN). Our experiments demonstrate that GVQA significantly outperforms SAN on both VQA-CP v1 and VQA-CP v2 datasets. Interestingly, it also outperforms more powerful VQA models such as Multimodal Compact Bilinear Pooling (MCB) in several cases. GVQA offers strengths complementary to SAN when trained and evaluated on the original VQA v1 and VQA v2 datasets. Finally, GVQA is more transparent and interpretable than existing VQA models.

****

Learning Pixel-Level Semantic Affinity With Image-Level Supervision for Weakly Supervised Semantic Segmentation

Jiwoon Ahn, Suha Kwak; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4981-4990

The deficiency of segmentation labels is one of the main obstacles to semantic segmentation in the wild. To alleviate this issue, we present a novel framework that generates segmentation labels of images given their image-level class labels. In this weakly supervised setting, trained models have been known to segment l ocal discriminative parts rather than the entire object area. Our solution is to propagate such local responses to nearby areas which belong to the same semantic entity. To this end, we propose a Deep Neural Network (DNN) called AffinityNet that predicts semantic affinity between a pair of adjacent image coordinates. The semantic propagation is then realized by random walk with the affinities predicted by AffinityNet. More importantly, the supervision employed to train AffinityNet is given by the initial discriminative part segmentation, which is incomplete as a segmentation annotation but sufficient for learning semantic affinities within small image areas. Thus the entire framework relies only on image-level class labels and does not require any extra data or annotations. On the PASCAL VOC 2012 dataset, a DNN learned with segmentation labels generated by our method outperforms previous models trained with the same level of supervision, and is even as competitive as those relying on stronger supervision.

****

From Lifestyle Vlogs to Everyday Interactions

David F. Fouhey, Wei-cheng Kuo, Alexei A. Efros, Jitendra Malik; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4991-5000

A major stumbling block to progress in understanding basic human interactions, such as getting out of bed or opening a refrigerator, is lack of good training data. Most past efforts have gathered this data explicitly: starting with a laundry list of action labels, and then querying search engines for videos tagged with each label. In this work, we do the reverse and search implicitly: we start with a large collection of interaction-rich video data and then annotate and analyze it. We use Internet Lifestyle Vlogs as the source of surprisingly large and diverse interaction data.   We show that by collecting the data first, we are able to achieve greater scale and far greater diversity in terms of actions and actors. Additionally, our data exposes biases built into common explicitly gathered data. We make sense of our data by analyzing the central component of interaction -- hands. We benchmark two tasks: identifying semantic object contact at the video level and non-semantic contact state at the frame level. We additionally demonstrate future prediction of hands.

****

Cross-Domain Weakly-Supervised Object Detection Through Progressive Domain Adaptation

Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, Kiyoharu Aizawa; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5001-5009

Can we detect common objects in a variety of image domains without instance-level annotations? In this paper, we present a framework for a novel task, cross-domain weakly supervised object detection, which addresses this question. For this paper, we have access to images with instance-level annotations in a source domain (e.g., natural image) and images with image-level annotations in a target domain (e.g., watercolor). In addition, the classes to be detected in the target domain are all or a subset of those in the source domain. Starting from a fully supervised object detector, which is pre-trained on the source domain, we propose a two-step progressive domain adaptation technique by fine-tuning the detector on two types of artificially and automatically generated samples. We test our methods on our newly collected datasets containing three image domains, and achieve an improvement of approximately 5 to 20 percentage points in terms of mean average precision (mAP) compared to the best-performing baselines.

****

RotationNet: Joint Object Categorization and Pose Estimation Using Multiviews From Unsupervised Viewpoints

Asako Kanezaki, Yasuyuki Matsushita, Yoshifumi Nishida; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5010-5019

We propose a Convolutional Neural Network (CNN)-based model ``RotationNet,'' which takes multi-view images of an object as input and jointly estimates its pose and object category. Unlike previous approaches that use known viewpoint labels for training, our method treats the viewpoint labels as latent variables, which are learned in an unsupervised manner during the training using an unaligned object dataset. RotationNet is designed to use only a partial set of multi-view images for inference, and this property makes it useful in practical scenarios where only partial views are available. Moreover, our pose alignment strategy enables one to obtain view-specific feature representations shared across classes, which is important to maintain high accuracy in both object categorization and pose estimation. Effectiveness of RotationNet is demonstrated by its superior performance to the state-of-the-art methods of 3D object classification on 10- and 40-class ModelNet datasets. We also show that RotationNet, even trained without known poses, achieves the state-of-the-art performance on an object pose estimation dataset.

****

An End-to-End TextSpotter With Explicit Alignment and Attention

Tong He, Zhi Tian, Weilin Huang, Chunhua Shen, Yu Qiao, Changming Sun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5020-5029

Text detection and recognition in natural images have long been considered as two separate tasks that are processed sequentially. Jointly training two tasks is non-trivial due to significant differences in learning difficulties and convergence rates. In this work, we present a conceptually simple yet efficient framework that simultaneously processes the two tasks in a united framework. Our main contributions are three-fold: (1) we propose a novel textalignment layer that allows it to precisely compute convolutional features of a text instance in arbitrary orientation, which is the key to boost the performance; (2) a character attention mechanism is introduced by using character spatial information as explicit supervision, leading to large improvements in recognition; (3) two technologies, together with a new RNN branch for word recognition, are integrated seamlessly into a single model which is end-to-end trainable. This allows the two tasks to work collaboratively by sharing convolutional features, which is critical to identify challenging text instances. Our model obtains impressive results in end-to-end recognition on the ICDAR 2015, significantly advancing the most recent results, with improvements of F-measure from (0.54, 0.51, 0.47) to (0.82, 0.77, 0.63), by using a strong, weak and generic lexicon respectively. Thanks to joint training, our method can also serve as a good detector by achieving a new state-of-the-art detection performance on related benchmarks. Code is available at https://github. com/tonghe90/textspotter.

****

WILDTRACK: A Multi-Camera HD Dataset for Dense Unscripted Pedestrian Detection

Tatjana Chavdarova, Pierre Baqué, Stéphane Bouquet, Andrii Maksai, Cijo Jose, Timur Bagautdinov, Louis Lettry, Pascal Fua, Luc Van Gool, François Fleuret; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5030-5039

People detection methods are highly sensitive to occlusions between pedestrians, which are extremely frequent in many situations where cameras have to be mounted at a limited height. The reduction of camera prices allows for the generalization of static multi-camera set-ups. Using joint visual information from multiple synchronized cameras gives the opportunity to improve detection performance. In this paper, we present a new large-scale and high-resolution dataset. It has been captured with seven static cameras in a public open area, and unscripted dense groups of pedestrians standing and walking. Together with the camera frames, we provide an accurate joint (extrinsic and intrinsic) calibration, as well as 7 series of 400 annotated frames for detection at a rate of 2 frames per second. This results in over 40,000 bounding boxes delimiting every person present in the area of interest, for a total of more than 300 individuals.   We provide a series of benchmark results using baseline algorithms published over the recent months for multi-view detection with deep neural networks, and trajectory estimation using a non-Markovian model.

****

Direct Shape Regression Networks for End-to-End Face Alignment

Xin Miao, Xiantong Zhen, Xianglong Liu, Cheng Deng, Vassilis Athitsos, Heng Huang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5040-5049

Face alignment has been extensively studied in computer vision community due to its fundamental role in facial analysis, but it remains an unsolved problem. The major challenges lie in the highly nonlinear relationship between face images and associated facial shapes, which is coupled by underlying correlation of landmarks. Existing methods mainly rely on cascaded regression, suffering from intrinsic shortcomings, e.g., strong dependency on initialization and failure to exploit landmark correlations. In this paper, we propose the direct shape regression network (DSRN) for end-to-end face alignment by jointly handling the aforementioned challenges in a unified framework. Specifically, by deploying doubly convolutional layer and by using the Fourier feature pooling layer proposed in this paper, DSRN efficiently constructs strong representations to disentangle highly nonlinear relationships between images and shapes; by incorporating a linear layer of low-rank learning, DSRN effectively encodes correlations of landmarks to improve performance. DSRN leverages the strengths of kernels for nonlinear feature extraction and neural networks for structured prediction, and provides the first end-to-end learning architecture for direct face alignment. Its effectiveness and generality are validated by extensive experiments on five benchmark datasets, including AFLW, 300W, CelebA, MAFL, and 300VW. All empirical results demonstrate that DSRN consistently produces high performance and in most cases surpasses state-of-the-art.

****

Natural and Effective Obfuscation by Head Inpainting

Qianru Sun, Liqian Ma, Seong Joon Oh, Luc Van Gool, Bernt Schiele, Mario Fritz; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5050-5059

As more and more personal photos are shared online, being able to obfuscate identities in such photos is becoming a necessity for privacy protection. People have largely resorted to blacking out or blurring head regions, but they result in poor user experience while being surprisingly ineffective against state of the art person recognizers[17]. In this work, we propose a novel head inpainting obfuscation technique. Generating a realistic head inpainting in social media photos is challenging because subjects appear in diverse activities and head orientations. We thus split the task into two sub-tasks: (1) facial landmark generation from image context (e.g. body pose) for seamless hypothesis of sensible head pose, and (2) facial landmark conditioned head inpainting. We verify that our inpainting method generates realistic person images, while achieving superior obfuscation performance against automatic person recognizers.

****

3D Semantic Trajectory Reconstruction From 3D Pixel Continuum

Jae Shin Yoon, Ziwei Li, Hyun Soo Park; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5060-5069

This paper presents a method to reconstruct dense semantic trajectory stream of human interactions in 3D from synchronized multiple videos. The interactions inherently introduce self-occlusion and illumination/appearance/shape changes, resulting in highly fragmented trajectory reconstruction with noisy and coarse semantic labels. Our conjecture is that among many views, there exists a set of views that can confidently recognize the visual semantic label of a 3D trajectory. We introduce a new representation called 3D semantic map---a probability distribution over the semantic labels per trajectory. We construct the 3D semantic map by reasoning about visibility and 2D recognition confidence based on view-pooling, i.e., finding the view that best represents the semantics of the trajectory. Using the 3D semantic map, we precisely infer all trajectory labels jointly by considering the affinity between long range trajectories via estimating their local rigid transformations. This inference quantitatively outperforms the baseline approaches in terms of predictive validity, representation robustness, and affinity effectiveness. We demonstrate that our algorithm can robustly compute the semantic labels of a large scale trajectory set involving real-world human interact ions with object, scenes, and people.

****

Optimizing Filter Size in Convolutional Neural Networks for Facial Action Unit Recognition

Shizhong Han, Zibo Meng, Zhiyuan Li, James O'Reilly, Jie Cai, Xiaofeng Wang, Yan Tong; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5070-5078

Recognizing facial action units (AUs) during spontaneous facial displays is a challenging problem. Most recently, Convolutional Neural Networks (CNNs) have shown promise for facial AU recognition, where predefined and fixed convolution filter sizes are employed. In order to achieve the best performance, the optimal filter size is often empirically found by conducting extensive experimental validation. Such a training process suffers from expensive training cost, especially as the network becomes deeper. This paper proposes a novel Optimized Filter Size CNN (OFS-CNN), where the filter sizes and weights of all convolutional layers are learned simultaneously from the training data along with learning convolution filters. Specifically, the filter size is defined as a continuous variable, which is optimized by minimizing the training loss. Experimental results on two AU-coded spontaneous databases have shown that the proposed OFS-CNN is capable of estimating optimal filter size for varying image resolution and outperforms traditional CNNs with the best filter size obtained by exhaustive search. The OFS-CNN also beats the CNN using multiple filter sizes and more importantly, is much more efficient during testing with the proposed forward-backward propagation algorithm.

****

V2V-PoseNet: Voxel-to-Voxel Prediction Network for Accurate 3D Hand and Human Pose Estimation From a Single Depth Map

Gyeongsik Moon, Ju Yong Chang, Kyoung Mu Lee; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5079-5088

Most of the existing deep learning-based methods for 3D hand and human pose estimation from a single depth map are based on a common framework that takes a 2D depth map and directly regresses the 3D coordinates of keypoints, such as hand or human body joints, via 2D convolutional neural networks (CNNs). The first weakness of this approach is the presence of perspective distortion in the 2D depth map. While the depth map is intrinsically 3D data, many previous methods treat depth maps as 2D images that can distort the shape of the actual object through projection from 3D to 2D space. This compels the network to perform perspective distortion-invariant estimation. The second weakness of the conventional approach is that directly regressing 3D coordinates from a 2D image is a highly non-linear mapping, which causes difficulty in the learning procedure. To overcome these weaknesses, we firstly cast the 3D hand and human pose estimation problem from a single depth map into a voxel-to-voxel prediction that uses a 3D voxelized grid and estimates the per-voxel likelihood for each keypoint. We design our model as a 3D CNN that provides accurate estimates while running in real-time. Our system outperforms previous methods in almost all publicly available 3D hand and human pose estimation datasets and placed first in the HANDS 2017 frame-based 3D hand pose estimation challenge. The code is available in https://github.com/mks0601/V2V-PoseNet_RELEASE.

****

Ring Loss: Convex Feature Normalization for Face Recognition

Yutong Zheng, Dipan K. Pal, Marios Savvides; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5089-5097

We motivate and present Ring loss, a simple and elegant feature normalization approach for deep networks designed to augment standard loss functions such as Softmax. We argue that deep feature normalization is an important aspect of supervised classification problems where we require the model to represent each class in a multi-class problem equally well. The direct approach to feature normalization through the hard normalization operation results in a non-convex formulation. Instead, Ring loss applies soft normalization, where it gradually learns to constrain the norm to the scaled unit circle while preserving convexity leading to

more robust features. We apply Ring loss to large-scale face recognition problems and present results on LFW, the challenging protocols of IJB-A Janus, Janus CS3 (a superset of IJB-A Janus), Celebrity Frontal-Profile (CFP) and MegaFace with 1 million distractors. Ring loss outperforms strong baselines, matches state-of-the-art performance on IJB-A Janus and outperforms all other results on the challenging Janus CS3 thereby achieving state-of-the-art. We also outperform strong baselines in handling extremely low resolution face matching.
****
Adversarially Occluded Samples for Person Re-Identification
Houjing Huang, Dangwei Li, Zhang Zhang, Xiaotang Chen, Kaiqi Huang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5098-5107
Person re-identification (ReID) is the task of retrieving particular persons across different cameras. Despite its great progress in recent years, it is still confronted with challenges like pose variation, occlusion, and similar appearance among different persons. The large gap between training and testing performance with existing models implies the insufficiency of generalization. Considering this fact, we propose to augment the variation of training data by introducing Adversarially Occluded Samples. These special samples are both a) meaningful in that they resemble real-scene occlusions, and b) effective in that they are tough for the original model and thus provide the momentum to jump out of local optimum. We mine these samples based on a trained ReID model and with the help of network visualization techniques. Extensive experiments show that the proposed samples help the model discover new discriminative clues on the body and generalize much better at test time. Our strategy makes significant improvement over strong baselines on three large-scale ReID datasets, Market1501, CUHK03 and DukeMTMC-reID.
****
Classifier Learning With Prior Probabilities for Facial Action Unit Recognition
Yong Zhang, Weiming Dong, Bao-Gang Hu, Qiang Ji; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5108-5116
Facial action units (AUs) play an important role in human emotion understanding. One big challenge for data-driven AU recognition approaches is the lack of enough AU annotations, since AU annotation requires strong domain expertise. To alleviate this issue, we propose a knowledge-driven method for jointly learning multiple AU classifiers without any AU annotation by leveraging prior probabilities on AUs, including expression-independent and expression-dependent AU probabilities. These prior probabilities are drawn from facial anatomy and emotion studies, and are independent of datasets. We incorporate the prior probabilities on AUs as the constraints into the objective function of multiple AU classifiers, and develop an efficient learning algorithm to solve the formulated problem. Experimental results on five benchmark expression databases demonstrate the effectiveness of the proposed method, especially its generalization ability, and the power of the prior probabilities.
****
4DFAB: A Large Scale 4D Database for Facial Expression Analysis and Biometric Applications
Shiyang Cheng, Irene Kotsia, Maja Pantic, Stefanos Zafeiriou; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5117-5126
The progress we are currently witnessing in many computer vision applications, including automatic face analysis, would not be made possible without tremendous efforts in collecting and annotating large scale visual databases. To this end, we propose 4DFAB, a new large scale database of dynamic high resolution 3D faces (over 1,800,000 3D meshes). 4DFAB contain recordings of 180 subjects captured in four different sessions spanned over a five-year period. It contains 4D videos of subjects displaying both spontaneous and posed facial behaviours. The database can be used for both face and facial expression recognition, as well as behavioural biometrics. It can also be used to learn very powerful blendshapes for parametrising facial behaviour. In this paper, we conduct several experiments and

demonstrate the usefulness of the database in various applications. The database will be made publicly available for research purposes.

****

Seeing Small Faces From Robust Anchor's Perspective

Chenchen Zhu, Ran Tao, Khoa Luu, Marios Savvides; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5127-5136

This paper introduces a novel anchor design principle to support anchor-based face detection for superior scale-invariant performance, especially on tiny faces. To achieve this, we explicitly address the problem that anchor-based detectors drop performance drastically on faces with tiny sizes, e.g. less than 16x16 pixels. In this paper, we investigate why this is the case. We discover that current anchor design cannot guarantee high overlaps between tiny faces and anchor boxes, which increases the difficulty of training. The new Expected Max Overlapping (EMO) score is proposed which can theoretically explain the low overlapping issue and inspire several effective strategies of new anchor design leading to higher face overlaps, including anchor stride reduction with new network architectures, extra shifted anchors, and stochastic face shifting. Comprehensive experiments show that our proposed method significantly outperforms the baseline anchor-based detector, while consistently achieving state-of-the-art results on challenging face detection datasets with competitive runtime speed.

****

2D/3D Pose Estimation and Action Recognition Using Multitask Deep Learning

Diogo C. Luvizon, David Picard, Hedi Tabia; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5137-5146

Action recognition and human pose estimation are closely related but both problems are generally handled as distinct tasks in the literature. In this work, we propose a multitask framework for jointly 2D and 3D pose estimation from still images and human action recognition from video sequences. We show that a single architecture can be used to solve the two problems in an efficient way and still achieves state-of-the-art results. Additionally, we demonstrate that optimization from end-to-end leads to significantly higher accuracy than separated learning. The proposed architecture can be trained with data from different categories simultaneously in a seamlessly way. The reported results on four datasets (MPII, Human3.6M, Penn Action and NTU) demonstrate the effectiveness of our method on the targeted tasks.

****

Dense 3D Regression for Hand Pose Estimation

Chengde Wan, Thomas Probst, Luc Van Gool, Angela Yao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5147-5156

We present a simple and effective method for 3D hand pose estimation from a single depth frame. As opposed to previous state-of-arts based on holistic 3D regression, our method works on dense pixel-wise estimation. This is achieved by careful design choices in pose parameterization, which leverages both 2D and 3D properties of depth map. Specifically, we decompose the pose parameters into a set of per-pixel estimations, i.e., 2D heat maps, 3D heat maps and unit 3D direction vector fields. The 2D/3D joint heat maps and 3D joint offsets are estimated via multi-task network cascades, which is trained end-to-end. The pixel-wise estimations can be directly translated into a vote casting scheme. A variant of mean shift is then used to aggregate local votes and explicitly handles the global 3D estimation in consensus with pixel-wise 2D and 3D estimations. Our method is efficient and highly accurate. On MSRA and NYU hand dataset, our method outperforms all previous state-of-arts by a large margin. On ICVL hand dataset, our method achieves similar accuracy compared to the state-of-art which is nearly saturated and outperforms other state-of-arts. Code will be made available.

****

Camera Style Adaptation for Person Re-Identification

Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, Yi Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5157-5166

Being a cross-camera retrieval task, person re-identification suffers from image style variations caused by different cameras. The art implicitly addresses this problem by learning a camera-invariant descriptor subspace. In this paper, we explicitly consider this challenge by introducing camera style (CamStyle) adaptation. CamStyle can serve as a data augmentation approach that smooths the camera style disparities. Specifically, with CycleGAN, labeled training images can be style-transferred to each camera, and, along with the original training samples, form the augmented training set. This method, while increasing data diversity against over-fitting, also incurs a considerable level of noise. In the effort to alleviate the impact of noise, the label smooth regularization (LSR) is adopted. The vanilla version of our method (without LSR) performs reasonably well on few-camera systems in which over-fitting often occurs. With LSR, we demonstrate consistent improvement in all systems regardless of the extent of over-fitting. We also report competitive accuracy compared with the state of the art.

****

PoseTrack: A Benchmark for Human Pose Estimation and Tracking

Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, Bernt Schiele; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5167-5176

Existing systems for video-based pose estimation and tracking struggle to perform well on realistic videos with multiple people and often fail to output body-pose trajectories consistent over time. To address this shortcoming this paper introduces PoseTrack which is a new large-scale benchmark for video-based human pose estimation and articulated tracking. Our new benchmark encompasses three tasks focusing on i) single-frame multi-person pose estimation, ii) multi-person pose estimation in videos, and iii) multi-person articulated tracking. To establish the benchmark, we collect, annotate and release a new dataset that features videos with multiple people labeled with person tracks and articulated pose. A public centralized evaluation server is provided to allow the research community to evaluate on a held-out test set. Furthermore, we conduct an extensive experimental study on recent approaches to articulated pose tracking and provide analysis of the strengths and weaknesses of the state of the art. We envision that the proposed benchmark will stimulate productive research both by providing a large and representative training dataset as well as providing a platform to objectively evaluate and compare the proposed methods. The benchmark is freely accessible at https://posetrack.net/.

****

Exploit the Unknown Gradually: One-Shot Video-Based Person Re-Identification by Stepwise Learning

Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, Yi Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5177-5186

We focus on the one-shot learning for video-based person re-Identification (re-ID). Unlabeled tracklets for the person re-ID tasks can be easily obtained by pre-processing, such as pedestrian detection and tracking. In this paper, we propose an approach to exploiting unlabeled tracklets by gradually but steadily improving the discriminative capability of the Convolutional Neural Network (CNN) feature representation via stepwise learning. We first initialize a CNN model using one labeled tracklet for each identity. Then we update the CNN model by the following two steps iteratively: 1. sample a few candidates with most reliable pseudo labels from unlabeled tracklets; 2. update the CNN model according to the selected data. Instead of the static sampling strategy applied in existing works, we propose a progressive sampling method to increase the number of the selected pseudo-labeled candidates step by step. We systematically investigate the way how we should select pseudo-labeled tracklets into the training set to make the best use of them. Notably, the rank-1 accuracy of our method outperforms the state-of-the-art method by 21.46 points (absolute, i.e., 62.67% vs. 41.21%) on the MARS dataset, and 16.53 points on the DukeMTMC-VideoReID dataset.

****

Pose-Robust Face Recognition via Deep Residual Equivariant Mapping

Kaidi Cao, Yu Rong, Cheng Li, Xiaoou Tang, Chen Change Loy; Proceedings of the I

EEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5187-5196

Face recognition achieves exceptional success thanks to the emergence of deep learning. However, many contemporary face recognition models still perform relatively poor in processing profile faces compared to frontal faces. A key reason is that the number of frontal and profile training faces are highly imbalanced - there are extensively more frontal training samples compared to profile ones. In addition, it is intrinsically hard to learn a deep representation that is geometrically invariant to large pose variations. In this study, we hypothesize that there is an inherent mapping between frontal and profile faces, and consequently, their discrepancy in the deep representation space can be bridged by an equivariant mapping. To exploit this mapping, we formulate a novel Deep Residual EquivAriant Mapping (DREAM) block, which is capable of adaptively adding residuals to the input deep representation to transform a profile face representation to a canonical pose that simplifies recognition. The DREAM block consistently enhances the performance of profile face recognition for many strong deep networks, including ResNet models, without deliberately augmenting training data of profile faces. The block is easy to use, light-weight, and can be implemented with a negligible computational overhead.

****

DecideNet: Counting Varying Density Crowds Through Attention Guided Detection and Density Estimation

Jiang Liu, Chenqiang Gao, Deyu Meng, Alexander G. Hauptmann; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5197-5206

In real-world crowd counting applications, the crowd densities vary greatly in spatial and temporal domains. A detection based counting method will estimate crowds accurately in low density scenes, while its reliability in congested areas is downgraded. A regression based approach, on the other hand, captures the general density information in crowded regions. Without knowing the location of each person, it tends to overestimate the count in low density areas. Thus, exclusively using either one of them is not sufficient to handle all kinds of scenes with varying densities. To address this issue, a novel end-to-end crowd counting framework, named DecideNet (DEteCtIon and Density Estimation Network) is proposed. It can adaptively decide the appropriate counting mode for different locations on the image based on its real density conditions. DecideNet starts with estimating the crowd density by generating detection and regression based density maps separately. To capture inevitable variation in densities, it incorporates an attention module, meant to adaptively assess the reliability of the two types of estimations. The final crowd counts are obtained with the guidance of the attention module to adopt suitable estimations from the two kinds of density maps. Experimental results show that our method achieves state-of-the-art performance on three challenging crowd counting datasets.

****

LSTM Pose Machines

Yue Luo, Jimmy Ren, Zhouxia Wang, Wenxiu Sun, Jinshan Pan, Jianbo Liu, Jiahao Pang, Liang Lin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5207-5215

We observed that recent state-of-the-art results on single image human pose estimation were achieved by multi-stage Convolution Neural Networks (CNN). Notwithstanding the superior performance on static images, the application of these models on videos is not only computationally intensive, it also suffers from performance degeneration and flicking. Such suboptimal results are mainly attributed to the inability of imposing sequential geometric consistency, handling severe image quality degradation (e.g. motion blur and occlusion) as well as the inability of capturing the temporal correlation among video frames. In this paper, we proposed a novel recurrent network to tackle these problems. We showed that if we were to impose the weight sharing scheme to the multi-stage CNN, it could be re-written as a Recurrent Neural Network (RNN). This property decouples the relationship among multiple network stages and results in significantly faster speed in i nvoking the network for videos. It also enables the adoption of Long Short-Term Memory (LSTM) units between video frames. We found such memory augmented RNN is very effective in imposing geometric consistency among frames. It also well handles input quality degradation in videos while successfully stabilizes the sequential outputs. The experiments showed that our approach significantly outperformed current state-of-the-art methods on two large-scale video pose estimation benchmarks. We also explored the memory cells inside the LSTM and provided insights on why such mechanism would benefit the prediction for video-based pose estimations.

****

Disentangling Features in 3D Face Shapes for Joint Face Reconstruction and Recognition

Feng Liu, Ronghang Zhu, Dan Zeng, Qijun Zhao, Xiaoming Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5216-5225

This paper proposes an encoder-decoder network to disentangle shape features during 3D face shape reconstruction from single 2D images, such that the tasks of learning discriminative shape features for face recognition and reconstructing accurate 3D face shapes can be done simultaneously. Unlike existing 3D face reconstruction methods, our proposed method directly regresses dense 3D face shapes from single 2D images, and tackles identity and residual (i.e., non-identity) components in 3D face shapes explicitly and separately based on a composite 3D face shape model with latent representations. We devise a training process for the proposed network with a joint loss measuring both face identification error and 3D face shape reconstruction error. We develop a multi image 3D morphable model (3DMM) fitting method for multiple 2D images of a subject to construct training data. Comprehensive experiments have been done on MICC, BU3DFE, LFW and YTF databases. The results show that our method expands the capacity of 3DMM for capturing discriminative shape features and facial detail, and thus outperforms existing methods both in 3D face reconstruction accuracy and in face recognition accuracy.

****

Convolutional Sequence to Sequence Model for Human Dynamics

Chen Li, Zhen Zhang, Wee Sun Lee, Gim Hee Lee; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5226-5234

Human motion modeling is a classic problem in com- puter vision and graphics. Challenges in modeling human motion include high dimensional prediction as well as extremely complicated dynamics.We present a novel approach to human motion modeling based on convolutional neural networks (CNN). The hierarchical structure of CNN makes it capable of capturing both spatial and temporal correlations effectively. In our proposed approach, a convolutional long-term encoder is used to encode the whole given motion sequence into a long-term hidden variable, which is used with a decoder to predict the remainder of the sequence. The decoder itself also has an encoder-decoder structure, in which the short-term encoder encodes a shorter sequence to a short-term hidden variable, and the spatial decoder maps the long and short-term hidden variable to motion predictions. By using such a model, we are able to capture both invariant and dynamic information of human motion, which results in more accurate predictions. Experiments show that our algorithm outperforms the state-of-the-art methods on the Human3.6M and CMU Motion Capture datasets. Our code is available at the project website

****

Gesture Recognition: Focus on the Hands

Pradyumna Narayana, Ross Beveridge, Bruce A. Draper; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5235-5244

Gestures are a common form of human communication and important for human computer interfaces (HCI). Recent approaches to gesture recognition use deep learning methods, including multi-channel methods. We show that when spatial channels are focused on the hands, gesture recognition improves significantly, particularly when the channels are fused using a sparse network. Using this technique, we improve performance on the ChaLearn IsoGD dataset from a previous best of 67.71% to

82.07%, and on the NVIDIA dataset from 83.8% to 91.28%.

****

Crowd Counting via Adversarial Cross-Scale Consistency Pursuit

Zan Shen, Yi Xu, Bingbing Ni, Minsi Wang, Jianguo Hu, Xiaokang Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5245-5254

Crowd counting or density estimation is a challenging task in computer vision due to large scale variations, perspective distortions and serious occlusions, etc. Existing methods generally suffers from two issues: 1) the model averaging effects in multi-scale CNNs induced by the widely adopted L2 regression loss; and 2) inconsistent estimation across different scaled inputs. To explicitly address these issues, we propose a novel crowd counting (density estimation) framework called Adversarial Cross-Scale Consistency Pursuit (ACSCP). On one hand, a U-net structural network is designed to generate density map from input patch, and an adversarial loss is employed to shrink the solution onto a realistic subspace, thus attenuating the blurry effects of density map estimation. On the other hand, we design a novel scale-consistency regularizer which enforces that the sum up of the crowd counts from local patches (i.e., small scale) is coherent with the overall count of their region union (i.e., large scale). The above losses are integrated via a joint training scheme, so as to help boost density estimation performance by further exploring the collaboration between both objectives. Extensive experiments on four benchmarks have well demonstrated the effectiveness of the proposed innovations as well as the superior performance over prior art.

****

3D Human Pose Estimation in the Wild by Adversarial Learning

Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, Xiaogang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5255-5264

Recently, remarkable advances have been achieved in 3D human pose estimation from monocular images because of the powerful Deep Convolutional Neural Networks (DCNNs). Despite their success on large-scale datasets collected in the constrained lab environment, it is difficult to obtain the 3D pose annotations for in-the-wild images. Therefore, 3D human pose estimation in the wild is still a challenge. In this paper, we propose an adversarial learning framework, which distills the 3D human pose structures learned from the fully annotated dataset to in-the-wild images with only 2D pose annotations. Instead of defining hard-coded rules to constrain the pose estimation results, we design a novel multi-source discriminator to distinguish the predicted 3D poses from the ground truth, which helps to enforce the pose estimator to generate anthropometrically valid poses even with images in the wild. We also observe that a carefully designed information source for the discriminator is essential to boost the performance. Thus, we design a geometric descriptor, which computes the pairwise relative locations and distances between body joints, as a new information source for the discriminator. The efficacy of our adversarial learning framework with the new geometric descriptor have been demonstrated through extensive experiments on two widely used public benchmarks. Our approach significantly improves the performance compared with previous state-of-the-art approaches.

****

CosFace: Large Margin Cosine Loss for Deep Face Recognition

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, Wei Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5265-5274

Face recognition has made extraordinary progress owing to the advancement of deep convolutional neural networks (CNNs). The central task of face recognition, including face verification and identification, involves face feature discrimination. However, the traditional softmax loss of deep CNNs usually lacks the power of discrimination. To address this problem, recently several loss functions such as center loss, large margin softmax loss, and angular softmax loss have been proposed. All these improved losses share the same idea: maximizing inter-class variance and minimizing intra-class variance. In this paper, we propose a novel lo ss function, namely large margin cosine loss (LMCL), to realize this idea from a different perspective. More specifically, we reformulate the softmax loss as a cosine loss by L2 normalizing both features and weight vectors to remove radial variations, based on which a cosine margin term is introduced to further maximize the decision margin in the angular space. As a result, minimum intra-class variance and maximum inter-class variance are achieved by virtue of normalization and cosine decision margin maximization. We refer to our model trained with LMCL as CosFace. Extensive experimental evaluations are conducted on the most popular public-domain face recognition datasets such as MegaFace Challenge, Youtube Faces (YTF) and Labeled Face in the Wild (LFW). We achieve the state-of-the-art performance on these benchmarks, which confirms the effectiveness of our proposed approach.

****

Encoding Crowd Interaction With Deep Neural Network for Pedestrian Trajectory Prediction

Yanyu Xu, Zhixin Piao, Shenghua Gao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5275-5284

Pedestrian trajectory prediction is a challenging task because of the complex nature of humans. In this paper, we tackle the problem within a deep learning framework by considering motion information of each pedestrian and its interaction with the crowd. Specifically, motivated by the residual learning in deep learning, we propose to predict displacement between neighboring frames for each pedestrian sequentially. To predict such displacement, we design a crowd interaction deep neural network (CIDNN) which considers the different importance of different pedestrians for the displacement prediction of a target pedestrian. Specifically, we use an LSTM to model motion information for all pedestrians and use a multi-layer perceptron to map the location of each pedestrian to a high dimensional feature space where the inner product between features is used as a measurement for the spatial affinity between two pedestrians. Then we weight the motion features of all pedestrians based on their spatial affinity to the target pedestrian for location displacement prediction. Extensive experiments on publicly available datasets validate the effectiveness of our method for trajectory prediction.

****

Mean-Variance Loss for Deep Age Estimation From a Face

Hongyu Pan, Hu Han, Shiguang Shan, Xilin Chen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5285-5294

Age estimation has broad application prospects of many fields, such as video surveillance, social networking, and human-computer interaction. However, many of the published age estimation approaches simply treat the age estimation as an exact age regression problem, and thus did not leverage a distribution's robustness in representing labels with ambiguity such as ages. In this paper, we propose a new loss function, called mean-variance loss, for robust age estimation via distribution learning. Specifically, the mean-variance loss consists of a mean loss, which penalizes difference between the mean of the estimated age distribution and the ground-truth age, and a variance loss, which penalizes the variance of the estimated age distribution to ensure a concentrated distribution. The proposed mean-variance loss and softmax loss are embedded jointly into Convolutional Neural Networks (CNNs) for age estimation, and the network weights are optimized via stochastic gradient descent (SGD) in an end-to-end learning way. Experimental results on a number of challenging face aging databases (FG-NET, MORPH Album II, and CLAP2016) show that the proposed approach outperforms the state-of-the-art methods by a large margin using a single model.

****

Probabilistic Joint Face-Skull Modelling for Facial Reconstruction

Dennis Madsen, Marcel Lüthi, Andreas Schneider, Thomas Vetter; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5295-5303

We present a novel method for co-registration of two independent statistical shape models. We solve the problem of aligning a face model to a skull model with stochastic optimization based on Markov Chain Monte Carlo (MCMC). We create a pro babilistic joint face-skull model and show how to obtain a distribution of plausible face shapes given a skull shape. Due to environmental and genetic factors, there exists a distribution of possible face shapes arising from the same skull. We pose facial reconstruction as a conditional distribution of plausible face shapes given a skull shape. Because it is very difficult to obtain the distribution directly from MRI or CT data, we create a dataset of artificial face-skull pairs. To do this, we propose to combine three data sources of independent origin to model the joint face-skull distribution: a face shape model, a skull shape model and tissue depth marker information. For a given skull, we compute the posterior distribution of faces matching the tissue depth distribution with Metropolis-Hastings. We estimate the joint face-skull distribution from samples of the posterior. To find faces matching to an unknown skull, we estimate the probability of the face under the joint face-skull model. To our knowledge, we are the first to provide a whole distribution of plausible faces arising from a skull instead of only a single reconstruction. We show how the face-skull model can be used to rank a face dataset and on average successfully identify the correct match in top 30%. The face ranking even works when obtaining the face shapes from 2D images. We furthermore show how the face-skull model can be useful to estimate the skull position in an MR-image.

****

Learning Latent Super-Events to Detect Multiple Activities in Videos

AJ Piergiovanni, Michael S. Ryoo; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5304-5313

In this paper, we introduce the concept of learning latent super-events from activity videos, and present how it benefits activity detection in continuous videos. We define a super-event as a set of multiple events occurring together in videos with a particular temporal organization; it is the opposite concept of sub-events. Real-world videos contain multiple activities and are rarely segmented (e.g., surveillance videos), and learning latent super-events allows the model to capture how the events are temporally related in videos. We design emph{temporal structure filters} that enable the model to focus on particular sub-intervals of the videos, and use them together with a soft attention mechanism to learn representations of latent super-events. Super-event representations are combined with per-frame or per-segment CNNs to provide frame-level annotations. Our approach is designed to be fully differentiable, enabling end-to-end learning of latent super-event representations jointly with the activity detector using them. Our experiments with multiple public video datasets confirm that the proposed concept of latent super-event learning significantly benefits activity detection, advancing the state-of-the-arts.

****

Temporal Hallucinating for Action Recognition With Few Still Images

Yali Wang, Lei Zhou, Yu Qiao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5314-5322

Action recognition in still images has been recently promoted by deep learning. However, the success of these deep models heavily depends on huge amount of training images for various action categories, which may not be available in practice. Alternatively, humans can classify new action categories after seeing few images, since we may not only compare appearance similarities between images on hand, but also attempt to recall importance motion cues from relevant action videos in our memory. To mimic this capacity, we propose a novel Hybrid Video Memory (HVM) machine, which can hallucinate temporal features of still images from video memory, in order to boost action recognition with few still images. First, we design a temporal memory module consisting of temporal hallucinating and predicting. Temporal hallucinating can generate temporal features of still images in an unsupervised manner. Hence, it can be flexibly used in realistic scenarios, where image and video categories may not be consistent. Temporal predicting can effectively infer action categories for query image, by integrating temporal features of training images and videos within a domain-adaptation manner. Second, we design a spatial memory module for spatial predicting. As spatial and temporal features are complementary to represent different actions, we apply spatial-tempora l prediction fusion to further boost performance. Finally, we design a video selection module to select strongly-relevant videos as memory. In this case, we can balance the number of images and videos to reduce prediction bias as well as preserve computation efficiency. To show the effectiveness, we conduct extensive experiments on three challenging data sets, where our HVM outperforms a number of recent approaches by temporal hallucinating from video memory.

****

Deep Progressive Reinforcement Learning for Skeleton-Based Action Recognition

Yansong Tang, Yi Tian, Jiwen Lu, Peiyang Li, Jie Zhou; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5323-5332

In this paper, we propose a deep progressive reinforcement learning (DPRL) method for action recognition in skeleton-based videos, which aims to distil the most informative frames and discard ambiguous frames in sequences for recognizing actions. Since the choices of selecting representative frames are multitudinous for each video, we model the frame selection as a progressive process through deep reinforcement learning, during which we progressively adjust the chosen frames by taking two important factors into account: (1) the quality of the selected frames and (2) the relationship between the selected frames to the whole video. Moreover, considering the topology of human body inherently lies in a graph-based structure, where the vertices and edges represent the hinged joints and rigid bones respectively, we employ the graph-based convolutional neural network to capture the dependency between the joints for action recognition. Our approach achieves very competitive performance on three widely used benchmarks.

****

Gaze Prediction in Dynamic 360° Immersive Videos

Yanyu Xu, Yanbing Dong, Junru Wu, Zhengzhong Sun, Zhiru Shi, Jingyi Yu, Shenghua Gao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5333-5342

This paper explores gaze prediction in dynamic 360^circ immersive videos, emph{i.e.}, based on the history scan path and VR contents, we predict where a viewer will look at an upcoming time. To tackle this problem, we first present the large-scale eye-tracking in dynamic VR scene dataset. Our dataset contains 208 360^circ videos captured in dynamic scenes, and each video is viewed by at least 31 subjects. Our analysis shows that gaze prediction depends on its history scan path and image contents. In terms of the image contents, those salient objects easily attract viewers' attention. On the one hand, the saliency is related to both appearance and motion of the objects. Considering that the saliency measured at different scales is different, we propose to compute saliency maps at different spatial scales: the sub-image patch centered at current gaze point, the sub-image corresponding to the Field of View (FoV), and the panorama image. Then we feed both the saliency maps and the corresponding images into a Convolutional Neural Network (CNN) for feature extraction. Meanwhile, we also use a Long-Short-Term-Memory (LSTM) to encode the history scan path. Then we combine the CNN features and LSTM features for gaze displacement prediction between gaze point at a current time and gaze point at an upcoming time. Extensive experiments validate the effectiveness of our method for gaze prediction in dynamic VR scenes.

****

When Will You Do What? - Anticipating Temporal Occurrences of Activities

Yazan Abu Farha, Alexander Richard, Juergen Gall; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5343-5352

Analyzing human actions in videos has gained increased attention recently. While most works focus on classifying and labeling observed video frames or anticipating the very recent future, making long-term predictions over more than just a few seconds is a task with many practical applications that has not yet been addressed. In this paper, we propose two methods to predict a considerably large amount of future actions and their durations. Both, a CNN and an RNN are trained to learn future video labels based on previously seen content. We show that our methods generate accurate predictions of the future even for long videos with a huge amount of different actions and can even deal with noisy or erroneous input information.

****

Fusing Crowd Density Maps and Visual Object Trackers for People Tracking in Crowd Scenes

Weihong Ren, Di Kang, Yandong Tang, Antoni B. Chan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5353-5362

While people tracking has been greatly improved over the recent years, crowd scenes remain particularly challenging for people tracking due to heavy occlusions, high crowd density, and significant appearance variation. To address these challenges, we first design a Sparse Kernelized Correlation Filter (S-KCF) to suppress target response variations caused by occlusions and illumination changes, and spurious responses due to similar distractor objects. We then propose a people tracking framework that fuses the S-KCF response map with an estimated crowd density map using a convolutional neural network (CNN), yielding a refined response map. To train the fusion CNN, we propose a two-stage strategy to gradually optimize the parameters. The first stage is to train a preliminary model in batch mode with image patches selected around the targets, and the second stage is to fine-tune the preliminary model using the real frame-by-frame tracking process. Our density fusion framework can significantly improves people tracking in crowd scenes, and can also be combined with other trackers to improve the tracking performance. We validate our framework on two crowd video datasets: UCSD and PETS2009.

****

Dual Attention Matching Network for Context-Aware Feature Sequence Based Person Re-Identification

Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C. Kot, Gang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5363-5372

Typical person re-identification (ReID) methods usually describe each pedestrian with a single feature vector and match them in a task-specific metric space. However, the methods based on a single feature vector are not sufficient enough to overcome visual ambiguity, which frequently occurs in real scenario. In this paper, we propose a novel end-to-end trainable framework, called Dual ATtention Matching network (DuATM), to learn context-aware feature sequences and perform attentive sequence comparison simultaneously. The core component of our DuATM framework is a dual attention mechanism, in which both intra-sequence and inter-sequence attention strategies are used for feature refinement and feature-pair alignment, respectively. Thus, detailed visual cues contained in the intermediate feature sequences can be automatically exploited and properly compared. We train the proposed DuATM network as a siamese network via a triplet loss assisted with a de-correlation loss and a cross-entropy loss. We conduct extensive experiments on both image and video based ReID benchmark datasets. Experimental results demonstrate the significant advantages of our approach compared to the state-of-the-art methods.

****

Easy Identification From Better Constraints: Multi-Shot Person Re-Identification From Reference Constraints

Jiahuan Zhou, Bing Su, Ying Wu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5373-5381

Multi-shot person re-identification (MsP-RID) utilizes multiple images from the same person to facilitate identification. Considering the fact that motion information may not be discriminative nor reliable enough for MsP-RID, this paper is focused on handling the large variations in the visual appearances through learning discriminative visual metrics for identification. Existing metric learning-based methods usually exploit pair-wise or triple-wise similarity constraints, that generally demands intensive optimization in metric learning, or leads to degraded performances by using sub-optimal solutions. In addition, as the training data are significantly imbalanced, the learning can be largely dominated by the negative pairs and thus produces unstable and non-discriminative results. In this paper, we propose a novel type of similarity constraint. It assigns the sample points to a set of ■extbf{reference points} to produce a linear number of ■extbf {reference constraints}. Several optimal transport-based schemes for reference constraint generation are proposed and studied. Based on those constraints, by utilizing a typical regressive metric learning model, the closed-form solution of the learned metric can be easily obtained. Extensive experiments and comparative studies on several public MsP-RID benchmarks have validated the effectiveness of our method and its significant superiority over the state-of-the-art MsP-RID methods in terms of both identification accuracy and running speed.

****

Crowd Counting With Deep Negative Correlation Learning

Zenglin Shi, Le Zhang, Yun Liu, Xiaofeng Cao, Yangdong Ye, Ming-Ming Cheng, Guoyan Zheng; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5382-5390

Deep convolutional networks (ConvNets) have achieved unprecedented performances on many computer vision tasks. However, their adaptations to crowd counting on single images are still in their infancy and suffer from severe over-fitting. Here we propose a new learning strategy to produce generalizable features by way of deep negative correlation learning (NCL). More specifically, we deeply learn a pool of decorrelated regressors with sound generalization capabilities through managing their intrinsic diversities. Our proposed method, named decorrelated ConvNet (D-ConvNet), is end-to-end-trainable and independent of the backbone fully-convolutional network architectures. Extensive experiments on very deep VGGNet as well as our customized network structure indicate the superiority of D-ConvNet when compared with several state-of-the-art methods. Our implementation will be released at https://github.com/shizenglin/Deep-NCL

****

Human Appearance Transfer

Mihai Zanfir, Alin-Ionut Popa, Andrei Zanfir, Cristian Sminchisescu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5391-5399

We propose an automatic person-to-person appearance transfer model based on explicit parametric 3d human representations and learned, constrained deep translation network architectures for photographic image synthesis. Given a single source image and a single target image, each corresponding to different human subjects, wearing different clothing and in different poses, our goal is to photo-realistically transfer the appearance from the source image onto the target image while preserving the target shape and clothing segmentation layout. Our solution to this new problem is formulated in terms of a computational pipeline that combines (1) 3d human pose and body shape estimation from monocular images, (2) identifying 3d surface colors elements (mesh triangles) visible in both images, that can be transferred directly using barycentric procedures, and (3) predicting surface appearance missing in the first image but visible in the second one using deep learning-based image synthesis techniques. Our model achieves promising results as supported by a perceptual user study where the participants rated around 65% of our results as good, very good or perfect, as well in automated tests (Inception scores and a Faster-RCNN human detector responding very similarly to real and model generated images). We further show how the proposed architecture can be profiled to automatically generate images of a person dressed with different clothing transferred from a person in another image, opening paths for applications in entertainment and photo-editing (e.g. embodying and posing as friends or famous actors), the fashion industry, or affordable online shopping of clothing.

****

Domain Generalization With Adversarial Feature Learning

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, Alex C. Kot; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5400-5409

In this paper, we tackle the problem of domain generalization: how to learn a generalized feature representation for an “unseen” target domain by taking the advantage of multiple seen source-domain data. We present a novel framework based on adversarial autoencoders to learn a generalized latent feature representation across domains for domain generalization. To be specific, we extend adversarial autoencoders by imposing the Maximum Mean Discrepancy (MMD) measure to align the distributions among different domains, and matching the aligned distribution to an arbitrary prior distribution via adversarial feature learning. In this way, the learned feature representation is supposed to be universal to the seen source domains because of the MMD regularization, and is expected to generalize well on the target domain because of the introduction of the prior distribution. We proposed an algorithm to jointly train different components of our proposed framework. Extensive experiments on various vision tasks demonstrate that our proposed framework can learn better generalized features for the unseen target domain compared with state of-the-art domain generalization methods.

****

Pyramid Stereo Matching Network

Jia-Ren Chang, Yong-Sheng Chen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5410-5418

Recent work has shown that depth estimation from a stereo pair of images can be formulated as a supervised learning task to be resolved with convolutional neural networks (CNNs). However, current architectures rely on patch-based Siamese networks, lacking the means to exploit context information for finding correspondence in ill-posed regions. To tackle this problem, we propose PSMNet, a pyramid stereo matching network consisting of two main modules: spatial pyramid pooling and 3D CNN. The spatial pyramid pooling module takes advantage of the capacity of global context information by aggregating context in different scales and locations to form a cost volume. The 3D CNN learns to regularize cost volume using stacked multiple hourglass networks in conjunction with intermediate supervision. The proposed approach was evaluated on several benchmark datasets. Our method ranked first in the KITTI 2012 and 2015 leaderboards before March 18, 2018. The codes of PSMNet are available at: https://github.com/JiaRenChang/PSMNet.

****

Event-Based Vision Meets Deep Learning on Steering Prediction for Self-Driving Cars

Ana I. Maqueda, Antonio Loquercio, Guillermo Gallego, Narciso García, Davide Scaramuzza; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5419-5427

Event cameras are bio-inspired vision sensors that naturally capture the dynamics of a scene, filtering out redundant information. This paper presents a deep neural network approach that unlocks the potential of event cameras on a challenging motion-estimation task: prediction of a vehicle’s steering angle. To make the best out of this sensor–algorithm combination, we adapt state-of-the-art convolutional architectures to the output of event sensors and extensively evaluate the performance of our approach on a publicly available large scale event-camera dataset (≈1000 km). We present qualitative and quantitative explanations of why event cameras allow robust steering prediction even in cases where traditional cameras fail, e.g. challenging illumination conditions and fast motion. Finally, we demonstrate the advantages of leveraging transfer learning from traditional to event-based vision, and show that our approach outperforms state-of-the-art algorithms based on standard cameras

****

Learning Answer Embeddings for Visual Question Answering

Hexiang Hu, Wei-Lun Chao, Fei Sha; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5428-5436

We propose a novel probabilistic model for visual question answering (Visual QA). The key idea is to infer two sets of embeddings: one for the image and the question jointly and the other for the answers. The learning objective is to learn the best parameterization of those embeddings such that the correct answer has higher likelihood among all possible answers. In contrast to several existing approaches of treating Visual QA as multi-way classification, the proposed approach takes the semantic relationships (as characterized by the embeddings) among answers into consideration, instead of viewing them as independent ordinal numbers. Thus, the learned embedded function can be used to embed unseen answers (in the training dataset). These properties make the approach particularly appealing f

or transfer learning for open-ended Visual QA, where the source dataset on which the model is learned has limited overlapping with the target dataset in the space of answers. We have also developed large-scale optimization techniques for applying the model to datasets with a large number of answers, where the challenge is to properly normalize the proposed probabilistic models. We validate our approach on several Visual QA datasets and investigate its utility for transferring models across datasets. The empirical results have shown that the approach performs well not only on in-domain learning but also on transfer learning.
****

Good View Hunting: Learning Photo Composition From Dense View Pairs

Zijun Wei, Jianming Zhang, Xiaohui Shen, Zhe Lin, Radomír Mech, Minh Hoai, Dimitris Samaras; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5437-5446

Finding views with good photo composition is a challenging task for machine learning methods. A key difficulty is the lack of well annotated large scale datasets. Most existing datasets only provide a limited number of annotations for good views, while ignoring the comparative nature of view selection. In this work, we present the first large scale Comparative Photo Composition dataset, which contains over one million comparative view pairs annotated using a cost-effective crowdsourcing workflow. We show that these comparative view annotations are essential for training a robust neural network model for composition. In addition, we propose a novel knowledge transfer framework to train a fast view proposal network, which runs at 75+ FPS and achieves state-of-the-art performance in image cropping and thumbnail generation tasks on three benchmark datasets. The superiority of our method is also demonstrated in a user study on a challenging experiment, where our method significantly outperforms the baseline methods in producing diversified well-composed views.
****

CleanNet: Transfer Learning for Scalable Image Classifier Training With Label Noise

Kuang-Huei Lee, Xiaodong He, Lei Zhang, Linjun Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5447-5456

In this paper, we study the problem of learning image classification models with label noise. Existing approaches depending on human supervision are generally not scalable as manually identifying correct or incorrect labels is time-consuming, whereas approaches not relying on human supervision are scalable but less effective. To reduce the amount of human supervision for label noise cleaning, we introduce CleanNet, a joint neural embedding network, which only requires a fraction of the classes being manually verified to provide the knowledge of label noise that can be transferred to other classes. We further integrate CleanNet and conventional convolutional neural network classifier into one framework for image classification learning. We demonstrate the effectiveness of the proposed algorithm on both of the label noise detection task and the image classification on noisy data task on several large-scale datasets. Experimental results show that CleanNet can reduce label noise detection error rate on held-out classes where no human supervision available by 41.5% compared to current weakly supervised methods. It also achieves 47% of the performance gain of verifying all images with only 3.2% images verified on an image classification task. Source code and dataset will be available at kuanghuei.github.io/CleanNetProject.
****

Independently Recurrent Neural Network (IndRNN): Building a Longer and Deeper RNN

Shuai Li, Wanqing Li, Chris Cook, Ce Zhu, Yanbo Gao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5457-5466

Recurrent neural networks (RNNs) have been widely used for processing sequential data. However, RNNs are commonly difficult to train due to the well-known gradient vanishing and exploding problems and hard to learn long-term patterns. Long short-term memory (LSTM) and gated recurrent unit (GRU) were developed to address these problems, but the use of hyperbolic tangent and the sigmoid action functions results in gradient decay over layers. Consequently, construction of an eff iciently trainable deep network is challenging. In addition, all the neurons in an RNN layer are entangled together and their behaviour is hard to interpret. To address these problems, a new type of RNN, referred to as independently recurrent neural network (IndRNN), is proposed in this paper, where neurons in the same layer are independent of each other and they are connected across layers. We have shown that an IndRNN can be easily regulated to prevent the gradient exploding and vanishing problems while allowing the network to learn long-term dependencies. Moreover, an IndRNN can work with non-saturated activation functions such as relu (rectified linear unit) and be still trained robustly. Multiple IndRNNs can be stacked to construct a network that is deeper than the existing RNNs. Experimental results have shown that the proposed IndRNN is able to process very long sequences (over 5000 time steps), can be used to construct very deep networks (21 layers used in the experiment) and still be trained robustly. Better performances have been achieved on various tasks by using IndRNNs compared with the traditional RNN and LSTM.

****

Mix and Match Networks: Encoder-Decoder Alignment for Zero-Pair Image Translation

Yaxing Wang, Joost van de Weijer, Luis Herranz; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5467-5476

We address the problem of image translation between domains or modalities for which no direct paired data is available (i.e. zero-pair translation). We propose mix and match networks, based on multiple encoders and decoders aligned in such a way that other encoder-decoder pairs can be composed at test time to perform unseen image translation tasks between domains or modalities for which explicit paired samples were not seen during training. We study the impact of autoencoders, side information and losses in improving the alignment and transferability of trained pairwise translation models to unseen translations. We show our approach is scalable and can perform colorization and style transfer between unseen combinations of domains. We evaluate our system in a challenging cross-modal setting where semantic segmentation is estimated from depth images, without explicit access to any depth-semantic segmentation training pairs. Our model outperforms baselines based on pix2pix and CycleGAN models.

****

Structured Uncertainty Prediction Networks

Garoe Dorta, Sara Vicente, Lourdes Agapito, Neill D. F. Campbell, Ivor Simpson; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5477-5485

This paper is the first work to propose a network to predict a structured uncertainty distribution for a synthesized image. Previous approaches have been mostly limited to predicting diagonal covariance matrices. Our novel model learns to predict a full Gaussian covariance matrix for each reconstruction, which permits efficient sampling and likelihood evaluation. We demonstrate that our model can accurately reconstruct ground truth correlated residual distributions for synthetic datasets and generate plausible high frequency samples for real face images. We also illustrate the use of these predicted covariances for structure preserving image denoising.

****

Between-Class Learning for Image Classification

Yuji Tokozume, Yoshitaka Ushiku, Tatsuya Harada; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5486-5494

In this paper, we propose a novel learning method for image classification called Between-Class learning (BC learning). We generate between-class images by mixing two images belonging to different classes with a random ratio. We then input the mixed image to the model and train the model to output the mixing ratio. BC learning has the ability to impose constraints on the shape of the feature distributions, and thus the generalization ability is improved. BC learning is originally a method developed for sounds, which can be digitally mixed. Mixing two image data does not appear to make sense; however, we argue that because convolutional neural networks have an aspect of treating input data as waveforms, what wor ks on sounds must also work on images. First, we propose a simple mixing method using internal divisions, which surprisingly proves to significantly improve performance. Second, we propose a mixing method that treats the images as waveforms, which leads to a further improvement in performance. As a result, we achieved 19.4% and 2.26% top-1 errors on ImageNet-1K and CIFAR-10, respectively.

****

Adversarial Feature Augmentation for Unsupervised Domain Adaptation

Riccardo Volpi, Pietro Morerio, Silvio Savarese, Vittorio Murino; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5495-5504

Recent works showed that Generative Adversarial Networks (GANs) can be successfully applied in unsupervised domain adaptation, where, given a labeled source dataset and an unlabeled target dataset, the goal is to train powerful classifiers for the target samples. In particular, it was shown that a GAN objective function can be used to learn target features indistinguishable from the source ones. In this work, we extend this framework by (i) forcing the learned feature extractor to be domain-invariant, and (ii) training it through data augmentation in the feature space, namely performing feature augmentation. While data augmentation in the image space is a well established technique in deep learning, feature augmentation has not yet received the same level of attention. We accomplish it by means of a feature generator trained by playing the GAN minimax game against source features. Results show that both enforcing domain-invariance and performing feature augmentation lead to superior or comparable performance to state-of-the-art results in several unsupervised domain adaptation benchmarks.

****

Generative Image Inpainting With Contextual Attention

Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, Thomas S. Huang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5505-5514

Recent deep learning based approaches have shown promising results for the challenging task of inpainting large missing regions in an image. These methods can generate visually plausible image structures and textures, but often create distorted structures or blurry textures inconsistent with surrounding areas. This is mainly due to ineffectiveness of convolutional neural networks in explicitly borrowing or copying information from distant spatial locations. On the other hand, traditional texture and patch synthesis approaches are particularly suitable when it needs to borrow textures from the surrounding regions. Motivated by these observations, we propose a new deep generative model-based approach which can not only synthesize novel image structures but also explicitly utilize surrounding image features as references during network training to make better predictions. The model is a feed-forward, fully convolutional neural network which can process images with multiple holes at arbitrary locations and with variable sizes during the test time. Experiments on multiple datasets including faces (CelebA, CelebA-HQ), textures (DTD) and natural images (ImageNet, Places2) demonstrate that our proposed approach generates higher-quality inpainting results than existing ones. Code, demo and models are available at: https://github.com/JiahuiYu/generative_inpainting.

****

CSGNet: Neural Shape Parser for Constructive Solid Geometry

Gopal Sharma, Rishabh Goyal, Difan Liu, Evangelos Kalogerakis, Subhransu Maji; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5515-5523

We present a neural architecture that takes as input a 2D or 3D shape and outputs a program that generates the shape. The instructions in our program are based on constructive solid geometry principles, i.e., a set of boolean operations on shape primitives defined recursively. Bottom-up techniques for this shape parsing task rely on primitive detection and are inherently slow since the search space over possible primitive combinations is large. In contrast, our model uses a recurrent neural network that parses the input shape in a top-down manner, which is significantly faster and yields a compact and easy-to-interpret sequence of modeling instructions. Our model is also more effective as a shape detector compared to existing state-of-the-art detection techniques. We finally demonstrate that our network can be trained on novel datasets without ground-truth program annotations through policy gradient techniques.

****

Conditional Image-to-Image Translation

Jianxin Lin, Yingce Xia, Tao Qin, Zhibo Chen, Tie-Yan Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5524-5532

Image-to-image translation tasks have been widely investigated with Generative Adversarial Networks (GANs) and dual learning. However, existing models lack the ability to control the translated results in the target domain and their results usually lack of diversity in the sense that a fixed image usually leads to (almost) deterministic translation result. In this paper, we study a new problem, conditional image-to-image translation, which is to translate an image from the source domain to the target domain conditioned on a given image in the target domain. It requires that the generated image should inherit some domain-specific features of the conditional image from the target domain. Therefore, changing the conditional image in the target domain will lead to diverse translation results for a fixed input image from the source domain, and therefore the conditional input image helps to control the translation results. We tackle this problem with unpaired data based on GANs and dual learning. We twist two conditional translation models (one translation from A domain to B domain, and the other one from B domain to A domain) together for inputs combination and reconstruction while preserving domain independent features. We carry out experiments on men's faces from-to women's faces translation and edges to shoes and bags translations. The results demonstrate the effectiveness of our proposed method.

****

Continuous Relaxation of MAP Inference: A Nonconvex Perspective

D. Khuê Lê-Huu, Nikos Paragios; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5533-5541

In this paper, we study a nonconvex continuous relaxation of MAP inference in discrete Markov random fields (MRFs). We show that for arbitrary MRFs, this relaxation is tight, and a discrete stationary point of it can be easily reached by a simple block coordinate descent algorithm. In addition, we study the resolution of this relaxation using popular gradient methods, and further propose a more effective solution using a multilinear decomposition framework based on the alternating direction method of multipliers (ADMM). Experiments on many real-world problems demonstrate that the proposed ADMM significantly outperforms other nonconvex relaxation based methods, and compares favorably with state of the art MRF optimization algorithms in different settings.

****

Feature Generating Networks for Zero-Shot Learning

Yongqin Xian, Tobias Lorenz, Bernt Schiele, Zeynep Akata; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5542-5551

Suffering from the extreme training data imbalance between seen and unseen classes, most of existing state-of-the-art approaches fail to achieve satisfactory results for the challenging generalized zero-shot learning task. To circumvent the need for labeled examples of unseen classes, we propose a novel generative adversarial network(GAN) that synthesizes CNN features conditioned on class-level semantic information, offering a shortcut directly from a semantic descriptor of a class to a class-conditional feature distribution. Our proposed approach, pairing a Wasserstein GAN with a classification loss, is able to generate sufficiently discriminative CNN features to train softmax classifiers or any multimodal embedding method. Our experimental results demonstrate a significant boost in accuracy over the state of the art on five challenging datasets -- CUB, FLO, SUN, AWA and ImageNet -- in both the zero-shot learning and generalized zero-shot learning settings.

****

Joint Optimization Framework for Learning With Noisy Labels

Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, Kiyoharu Aizawa; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5552-5560

Deep neural networks (DNNs) trained on large-scale datasets have exhibited significant performance in image classification. Many large-scale datasets are collected from websites, however they tend to contain inaccurate labels that are termed as noisy labels. Training on such noisy labeled datasets causes performance degradation because DNNs easily overfit to noisy labels. To overcome this problem, we propose a joint optimization framework of learning DNN parameters and estimating true labels. Our framework can correct labels during training by alternating update of network parameters and labels. We conduct experiments on the noisy CIFAR-10 datasets and the Clothing1M dataset. The results indicate that our approach significantly outperforms other state-of-the-art methods.

****

Convolutional Image Captioning

Jyoti Aneja, Aditya Deshpande, Alexander G. Schwing; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5561-5570

Image captioning is an important task, applicable to virtual assistants, editing tools, image indexing, and support of the disabled. In recent years significant progress has been made in image captioning, using Recurrent Neural Networks powered by long-short term-memory (LSTM) units. Despite mitigating the vanishing gradient problem, and despite their compelling ability to memorize dependencies, LSTM units are complex and inherently sequential across time. To address this issue, recent work has shown benefits of convolutional networks for machine translation and conditional image generation. Inspired by their success, in this paper, we develop a convolutional image captioning technique. We demonstrate its efficacy on the challenging MSCOCO dataset and demonstrate performance on par with the LSTM baseline, while having a faster training time per number of parameters. We also perform a detailed analysis, providing compelling reasons in favor of convolutional language generation approaches.

****

AON: Towards Arbitrarily-Oriented Text Recognition

Zhanzhan Cheng, Yangliu Xu, Fan Bai, Yi Niu, Shiliang Pu, Shuigeng Zhou; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5571-5579

Recognizing text from natural images is a hot research topic in computer vision due to its various applications. Despite the enduring research of several decades on optical character recognition (OCR), recognizing texts from natural images is still a challenging task. This is because scene texts are often in irregular (e.g. curved, arbitrarily-oriented or seriously distorted) arrangements, which have not yet been well addressed in the literature. Existing methods on text recognition mainly work with regular (horizontal and frontal) texts and cannot be trivially generalized to handle irregular texts. In this paper, we develop the arbitrary orientation network (AON) to directly capture the deep features of irregular texts, which are combined into an attention-based decoder to generate character sequence. The whole network can be trained end-to-end by using only images and word-level annotations. Extensive experiments on various benchmarks, including the CUTE80, SVT-Perspective, IIIT5k, SVT and ICDAR datasets, show that the proposed AON-based method achieves the-state-of-the-art performance in irregular datasets, and is comparable to major existing methods in regular datasets.

****

Wrapped Gaussian Process Regression on Riemannian Manifolds

Anton Mallasto, Aasa Feragen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5580-5588

Gaussian process (GP) regression is a powerful tool in non-parametric regression providing uncertainty estimates. However, it is limited to data in vector spaces. In fields such as shape analysis and diffusion tensor imaging, the data often lies on a manifold, making GP regression non- viable, as the resulting predictive distribution does not live in the correct geometric space. We tackle the prob lem by defining wrapped Gaussian processes (WGPs) on Rieman- nian manifolds, using the probabilistic setting to general- ize GP regression to the context of manifold-valued targets. The method is validated empirically on diffusion weighted imaging (DWI) data, directional data on the sphere and in the Kendall shape space, endorsing WGP regression as an efficient and flexible tool for manifold-valued regression.

****

Geometry Guided Convolutional Neural Networks for Self-Supervised Video Representation Learning

Chuang Gan, Boqing Gong, Kun Liu, Hao Su, Leonidas J. Guibas; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5589-5597

It is often laborious and costly to manually annotate videos for training high-quality video recognition models, so there has been some work and interest in exploring alternative, cheap, and yet often noisy and indirect, training signals for learning the video representations. However, these signals are still coarse, supplying supervision at the whole video frame level, and subtle, sometimes enforcing the learning agent to solve problems that are even hard for humans. In this paper, we instead explore geometry, a grand new type of auxiliary supervision for the self-supervised learning of video representations. In particular, we extract pixel-wise geometry information as flow fields and disparity maps from synthetic imagery and real 3D movies. Although the geometry and high-level semantics are seemingly distant topics, surprisingly, we find that the convolutional neural networks pre-trained by the geometry cues can be effectively adapted to semantic video understanding tasks. In addition, we also find that a progressive training strategy can foster a better neural network for the video recognition task than blindly pooling the distinct sources of geometry cues together. Extensive results on video dynamic scene recognition and action recognition tasks show that our geometry guided networks significantly outperform the competing methods that are trained with other types of labeling-free supervision signals.

****

DiverseNet: When One Right Answer Is Not Enough

Michael Firman, Neill D. F. Campbell, Lourdes Agapito, Gabriel J. Brostow; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5598-5607

Many structured prediction tasks in machine vision have a collection of acceptable answers, instead of one definitive ground truth answer. Segmentation of images, for example, is subject to human labeling bias. Similarly, there are multiple possible pixel values that could plausibly complete occluded image regions. State-of-the art supervised learning methods are typically optimized to make a single test-time prediction for each query, failing to find other modes in the output space. Existing methods that allow for sampling often sacrifice speed or accuracy. We introduce a simple method for training a neural network, which enables diverse structured predictions to be made for each test-time query. For a single input, we learn to predict a range of possible answers. We compare favorably to methods that seek diversity through an ensemble of networks. Such stochastic multiple choice learning faces mode collapse, where one or more ensemble members fail to receive any training signal. Our best performing solution can be deployed for various tasks, and just involves small modifications to the existing single-mode architecture, loss function, and training regime. We demonstrate that our method results in quantitative improvements across three challenging tasks: 2D image completion, 3D volume estimation, and flow prediction.

****

Deep Face Detector Adaptation Without Negative Transfer or Catastrophic Forgetting

Muhammad Abdullah Jamal, Haoxiang Li, Boqing Gong; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5608-5618

Arguably, no single face detector fits all real-life scenarios. It is often desirable to have some built-in schemes for a face detector to automatically adapt, e.g., to a particular user's photo album (the target domain). We propose a novel face detector adaptation approach that works as long as there are representative images of the target domain no matter they are labeled or not and, more importantly, without the need of accessing the training data of the source domain. Our approach explicitly accounts for the notorious negative transfer caveat in domain adaptation thanks to a residual loss by design. Moreover, it does not incur catastrophic interference with the knowledge learned from the source domain and, therefore, the adapted face detectors maintain about the same performance as the old detectors in the original source domain. As such, our adaption approach to face detectors is analogous to the popular interpolation techniques for language models; it may opens a new direction for progressively training the face detectors domain by domain. We report extensive experimental results to verify our approach on two massively benchmarked face detectors.

****

Analyzing Filters Toward Efficient ConvNet

Takumi Kobayashi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5619-5628

Deep convolutional neural network (ConvNet) is a promising approach for high-performance image classification. The behavior of ConvNet is analyzed mainly based on the neuron activations, such as by visualizing them. In this paper, in contrast to the activations, we focus on filters which are main components of ConvNets. Through analyzing two types of filters at convolution and fully-connected (FC) layers, respectively, on various pre-trained ConvNets, we present the methods to efficiently reformulate the filters, contributing to improving both memory size and classification performance of the ConvNets. They render the filter bases formulated in a parameter-free form as well as the efficient representation for the FC layer. The experimental results on image classification show that the methods are favorably applied to improve various ConvNets, including ResNet, trained on ImageNet with exhibiting high transferability on the other datasets.

****

Regularizing Deep Networks by Modeling and Predicting Label Structure

Mohammadreza Mostajabi, Michael Maire, Gregory Shakhnarovich; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5629-5638

We construct custom regularization functions for use in supervised training of deep neural networks. Our technique is applicable when the ground-truth labels themselves exhibit internal structure; we derive a regularizer by learning an autoencoder over the set of annotations. Training thereby becomes a two-phase procedure. The first phase models labels with an autoencoder. The second phase trains the actual network of interest by attaching an auxiliary branch that must predict output via a hidden layer of the autoencoder. After training, we discard this auxiliary branch. We experiment in the context of semantic segmentation, demonstrating this regularization strategy leads to consistent accuracy boosts over baselines, both when training from scratch, or in combination with ImageNet pretraining. Gains are also consistent over different choices of convolutional network architecture. As our regularizer is discarded after training, our method has zero cost at test time; the performance improvements are essentially free. We are simply able to learn better network weights by building an abstract model of the label space, and then training the network to understand this abstraction alongside the original task.

****

In-Place Activated BatchNorm for Memory-Optimized Training of DNNs

Samuel Rota Bulò, Lorenzo Porzi, Peter Kontschieder; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5639-5647

In this work we present In-Place Activated Batch Normalization (InPlace-ABN) -- a novel approach to drastically reduce the training memory footprint of modern deep neural networks in a computationally efficient way. Our solution substitutes the conventionally used succession of BatchNorm + Activation layers with a single plugin layer, hence avoiding invasive framework surgery while providing straightforward applicability for existing deep learning frameworks. We obtain memory savings of up to 50% by dropping intermediate results and by recovering require d information during the backward pass through the inversion of stored forward results, with only minor increase (0.8-2%) in computation time. Also, we demonstrate how frequently used checkpointing approaches can be made computationally as efficient as InPlace-ABN. In our experiments on image classification, we demonstrate on-par results on ImageNet-1k with state-of-the-art approaches. On the memory-demanding task of semantic segmentation, we report competitive results for COCO-Stuff and set new state-of-the-art results for Cityscapes and Mapillary Vistas. Code can be found at https://github.com/mapillary/inplace_abn.

****

DVQA: Understanding Data Visualizations via Question Answering

Kushal Kafle, Brian Price, Scott Cohen, Christopher Kanan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5648-5656

Bar charts are an effective way to convey numeric information, but today's algorithms cannot parse them. Existing methods fail when faced with even minor variations in appearance. Here, we present DVQA, a dataset that tests many aspects of bar chart understanding in a question answering framework. Unlike visual question answering (VQA), DVQA requires processing words and answers that are unique to a particular bar chart. State-of-the-art VQA algorithms perform poorly on DVQA, and we propose two strong baselines that perform considerably better. Our work will enable algorithms to automatically extract numeric and semantic information from vast quantities of bar charts found in scientific publications, Internet articles, business reports, and many other areas.

****

DA-GAN: Instance-Level Image Translation by Deep Attention Generative Adversarial Networks

Shuang Ma, Jianlong Fu, Chang Wen Chen, Tao Mei; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5657-5666

Unsupervised image translation, which aims in translating two independent sets of images, is challenging in discovering the correct correspondences without paired data. Existing works build upon Generative Adversarial Networks (GANs) such that the distribution of the translated images are indistinguishable from the distribution of the target set. However, such set-level constraints cannot learn the instance-level correspondences (e.g. aligned semantic parts in object transfiguration task). This limitation often results in false positives (e.g. geometric or semantic artifacts), and further leads to mode collapse problem. To address the above issues, we propose a novel framework for instance-level image translation by Deep Attention GAN (DA-GAN). Such a design enables DA-GAN to decompose the task of translating samples from two sets into translating instances in a highly-structured latent space. Specifically, we jointly learn a deep attention encoder, and the instance-level correspondences could be consequently discovered through attending on the learned instances. Therefore, the constraints could be exploited on both set-level and instance-level. Comparisons against several state-of-the- arts demonstrate the superiority of our approach, and the broad application capability, e.g, pose morphing, data augmentation, etc., pushes the margin of domain translation problem.

****

Unsupervised Learning of Depth and Ego-Motion From Monocular Video Using 3D Geometric Constraints

Reza Mahjourian, Martin Wicke, Anelia Angelova; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5667-5675

We present a novel approach for unsupervised learning of depth and ego-motion from monocular video. Unsupervised learning removes the need for separate supervisory signals (depth or ego-motion ground truth, or multi-view video). Prior work in unsupervised depth learning uses pixel-wise or gradient-based losses, which only consider pixels in small local neighborhoods. Our main contribution is to explicitly consider the inferred 3D geometry of the whole scene, and enforce consistency of the estimated 3D point clouds and ego-motion across consecutive frames. This is a challenging task and is solved by a novel (approximate) backpropagation algorithm for aligning 3D structures. We combine this novel 3D-based loss

with 2D losses based on photometric quality of frame reconstructions using estimated depth and ego-motion from adjacent frames. We also incorporate validity masks to avoid penalizing areas in which no useful information exists. We test our algorithm on the KITTI dataset and on a video dataset captured on an uncalibrated mobile phone camera. Our proposed approach consistently improves depth estimates on both datasets, and outperforms the state-of-the-art for both depth and ego-motion. Because we only require a simple video, learning depth and ego-motion on large and varied datasets becomes possible. We demonstrate this by training on the low quality uncalibrated video dataset and evaluating on KITTI, ranking among top performing prior methods which are trained on KITTI itself.
****

FOTS: Fast Oriented Text Spotting With a Unified Network

Xuebo Liu, Ding Liang, Shi Yan, Dagui Chen, Yu Qiao, Junjie Yan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5676-5685

Incidental scene text spotting is considered one of the most difficult and valuable challenges in the document analysis community. Most existing methods treat text detection and recognition as separate tasks. In this work, we propose a unified end-to-end trainable Fast Oriented Text Spotting (FOTS) network for simultaneous detection and recognition, sharing computation and visual information among the two complementary tasks. Specifically, RoIRotate is introduced to share convolutional features between detection and recognition. Benefiting from convolution sharing strategy, our FOTS has little computation overhead compared to baseline text detection network, and the joint training method makes our method perform better than these two-stage methods. Experiments on ICDAR 2015, ICDAR 2017 MLT, and ICDAR 2013 datasets demonstrate that the proposed method outperforms state-of-the-art methods significantly, which further allows us to develop the first real-time oriented text spotting system which surpasses all previous state-of-the-art results by more than 5% on ICDAR 2015 text spotting task while keeping 22.6 fps.
****

Mobile Video Object Detection With Temporally-Aware Feature Maps

Mason Liu, Menglong Zhu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5686-5695

This paper introduces an online model for object detection in videos with real-time performance on mobile and embedded devices. Our approach combines fast single-image object detection with convolutional long short term memory (LSTM) layers to create an interweaved recurrent-convolutional architecture. Additionally, we propose an efficient Bottleneck-LSTM layer that significantly reduces computational cost compared to regular LSTMs. Our network achieves temporal awareness by using Bottleneck-LSTMs to refine and propagate feature maps across frames. This approach is substantially faster than existing detection methods in video, outperforming the fastest single-frame models in model size and computational cost while attaining accuracy comparable to much more expensive single-frame models on the Imagenet VID 2015 dataset. Our model reaches a real-time inference speed of up to 15 FPS on a mobile CPU.
****

Weakly Supervised Phrase Localization With Multi-Scale Anchored Transformer Network

Fang Zhao, Jianshu Li, Jian Zhao, Jiashi Feng; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5696-5705

In this paper, we propose a novel weakly supervised model, Multi-scale Anchored Transformer Network (MATN), to accurately localize free-form textual phrases with only image-level supervision. The proposed MATN takes region proposals as localization anchors, and learns a multi-scale correspondence network to continuously search for phrase regions referring to the anchors. In this way, MATN can exploit useful cues from these anchors to reliably reason about locations of the regions described by the phrases given only image-level supervision. Through differentiable sampling on image spatial feature maps, MATN introduces a novel training objective to simultaneously minimize a contrastive reconstruction loss between different phrases from a single image and a set of triplet losses among multiple images with similar phrases. Superior to existing region proposal based methods, MATN searches for the optimal bounding box over the entire feature map instead of selecting a sub-optimal one from discrete region proposals. We evaluate MATN on the Flickr30K Entities and ReferItGame datasets. The experimental results show that MATN significantly outperforms the state-of-the-art methods.

****

Revisiting Oxford and Paris: Large-Scale Image Retrieval Benchmarking

Filip Radenovi■, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, Ond■ej Chum; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5706-5715

In this paper we address issues with image retrieval benchmarking on standard and popular Oxford 5k and Paris 6k datasets. In particular, annotation errors, the size of the dataset, and the level of challenge are addressed: new annotation for both datasets is created with an extra attention to the reliability of the ground truth. Three new protocols of varying difficulty are introduced. The protocols allow fair comparison between different methods, including those using a dataset pre-processing stage. For each dataset, 15 new challenging queries are introduced. Finally, a new set of 1M hard, semi-automatically cleaned distractors is selected. An extensive comparison of the state-of-the-art methods is performed on the new benchmark. Different types of methods are evaluated, ranging from local-feature-based to modern CNN based methods. The best results are achieved by taking the best of the two worlds. Most importantly, image retrieval appears far from being solved.

****

Cross-Dataset Adaptation for Visual Question Answering

Wei-Lun Chao, Hexiang Hu, Fei Sha; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5716-5725

We investigate the problem of cross-dataset adaptation for visual question answering (Visual QA). Our goal is to train a Visual QA model on a source dataset but apply it to another target one. Analogous to domain adaptation for visual recognition, this setting is appealing when the target dataset does not have a sufficient amount of labeled data to learn an ``in-domain'' model. The key challenge is that the two datasets are constructed differently, resulting in the cross-dataset mismatch on images, questions, or answers. We overcome this difficulty by proposing a novel domain adaptation algorithm. Our method reduces the difference in statistical distributions by transforming the feature representation of the data in the target dataset. Moreover, it maximizes the likelihood of answering questions (in the target dataset) correctly using the Visual QA model trained on the source dataset. We empirically studied the effectiveness of the proposed approach on adapting among several popular Visual QA datasets. We show that the proposed method improves over baselines where there is no adaptation and several other adaptation methods. We both quantitatively and qualitatively analyze when the adaptation can be mostly effective.

****

Globally Optimal Inlier Set Maximization for Atlanta Frame Estimation

Kyungdon Joo, Tae-Hyun Oh, In So Kweon, Jean-Charles Bazin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5726-5734

In this work, we describe man-made structures via an appropriate structure assumption, called Atlanta world, which contains a vertical direction (typically the gravity direction) and a set of horizontal directions orthogonal to the vertical direction. Contrary to the commonly used Manhattan world assumption, the horizontal directions in Atlanta world are not necessarily orthogonal to each other. While Atlanta world permits to encompass a wider range of scenes, this makes the solution space larger and the problem more challenging. Given a set of inputs, such as lines in a calibrated image or surface normals, we propose the first globally optimal method of inlier set maximization for Atlanta direction estimation. We define a novel search space for Atlanta world, as well as its parameterization, and solve this challenging problem by a branch-and-bound framework. Experime

ntal results with synthetic and real-world datasets have successfully confirmed the validity of our approach.
****

End-to-End Convolutional Semantic Embeddings

Quanzeng You, Zhengyou Zhang, Jiebo Luo; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5735-5744

Semantic embeddings for images and sentences have been widely studied recently. The ability of deep neural networks on learning rich and robust visual and textual representations offers the opportunity to develop effective semantic embedding models. Currently, the state-of-the-art approaches in semantic learning first employ deep neural networks to encode images and sentences into a common semantic space. Then, the learning objective is to ensure a larger similarity between matching image and sentence pairs than randomly sampled pairs. Usually, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are employed for learning image and sentence representations, respectively. On one hand, CNNs are known to produce robust visual features at different levels and RNNs are known for capturing dependencies in sequential data. Therefore, this simple framework can be sufficiently effective in learning visual and textual semantics. On the other hand, different from CNNs, RNNs cannot produce middle-level (e.g. phrase-level in text) representations. As a result, only global representations are available for semantic learning. This could potentially limit the performance of the model due to the hierarchical structures in images and sentences. In this work, we apply Convolutional Neural Networks to process both images and sentences. Consequently, we can employ mid-level representations to assist global semantic learning by introducing a new learning objective on the convolutional layers. The experimental results show that our proposed textual CNN models with the new learning objective lead to better performance than the state-of-the-art approaches.
****

Referring Image Segmentation via Recurrent Refinement Networks

Ruiyu Li, Kaican Li, Yi-Chun Kuo, Michelle Shu, Xiaojuan Qi, Xiaoyong Shen, Jiaya Jia; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5745-5753

We address the problem of image segmentation from natural language descriptions. Existing deep learning-based methods encode image representations based on the output of the last convolutional layer. One general issue is that the resulting image representation lacks multi-scale semantics, which are key components in advanced segmentation systems. In this paper, we utilize the feature pyramids inherently existing in convolutional neural networks to capture the semantics at different scales. To produce suitable information flow through the path of feature hierarchy, we propose Recurrent Refinement Network (RRN) that takes pyramidal features as input to refine the segmentation mask progressively. Experimental results on four available datasets show that our approach outperforms multiple baselines and state-of-the-art.
****

Two Can Play This Game: Visual Dialog With Discriminative Question Generation and Answering

Unnat Jain, Svetlana Lazebnik, Alexander G. Schwing; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5754-5763

Human conversation is a complex mechanism with subtle nuances. It is hence an ambitious goal to develop artificial intelligence agents that can participate fluently in a conversation. While we are still far from achieving this goal, recent progress in visual question answering, image captioning, and visual question generation shows that dialog systems may be realizable in the not too distant future. To this end, a novel dataset was introduced recently and encouraging results were demonstrated, particularly for question answering. In this paper, we demonstrate a simple symmetric discriminative baseline, that can be applied to both predicting an answer as well as predicting a question. We show that this method performs on par with the state of the art, even memory net based methods. In addition, for the first time on the visual dialog dataset, we assess the performance of a system asking questions, and demonstrate how visual dialog can be generated from discriminative question generation and question answering.

****

Generative Adversarial Learning Towards Fast Weakly Supervised Detection

Yunhan Shen, Rongrong Ji, Shengchuan Zhang, Wangmeng Zuo, Yan Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5764-5773

Weakly supervised object detection has attracted extensive research efforts in recent years. Without the need of annotating bounding boxes, the existing methods usually follow a two/multi-stage pipeline with an online compulsive stage to extract object proposals, which is an order of magnitude slower than fast fully supervised object detectors such as SSD [31] and YOLO [34]. In this paper, we speedup online weakly supervised object detectors by orders of magnitude by proposing a novel generative adversarial learning paradigm. In the proposed paradigm, the generator is a one-stage object detector to generate bounding boxes from images. To guide the learning of object-level generator, a surrogator is introduced to mine high-quality bounding boxes for training. We further adapt a structural similarity loss in combination with an adversarial loss into the training objective, which solves the challenge that the bounding boxes produced by the surrogator may not well capture their ground truth. Our one-stage detector outperforms all existing schemes in terms of detection accuracy, running at 118 frames per second, which is up to 438x faster than the state-of-the-art weakly supervised detectors [8, 30, 15, 27, 45]. The code will be available publicly soon.

****

A Deeper Look at Power Normalizations

Piotr Koniusz, Hongguang Zhang, Fatih Porikli; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5774-5783

Power Normalizations (PN) are very useful non-linear operators in the context of Bag-of-Words data representations as they tackle problems such as feature imbalance. In this paper, we reconsider these operators in the deep learning setup by introducing a novel layer that implements PN for non-linear pooling of feature maps. Specifically, by using a kernel formulation, our layer combines the feature vectors and their respective spatial locations in the feature maps produced by the last convolutional layer of CNN. Linearization of such a kernel results in a positive definite matrix capturing the second-order statistics of the feature vectors, to which PN operators are applied. We study two types of PN functions, namely (i) MaxExp and (ii) Gamma, addressing their role and meaning in the context of non-linear pooling. We also provide a probabilistic interpretation of these operators and derive their surrogates with well-behaved gradients for end-to-end CNN learning. We apply our theory to practice by implementing the PN layer on a ResNet-50 model and showcase experiments on four benchmarks for fine-grained recognition, scene recognition, and material classification. Our results demonstrate state-of-the-part performance across all these tasks.

****

Dimensionality's Blessing: Clustering Images by Underlying Distribution

Wen-Yan Lin, Siying Liu, Jian-Huang Lai, Yasuyuki Matsushita; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5784-5793

Many high dimensional vector distances tend to a constant. This is typically considered a negative “contrast-loss” phenomenon that hinders clustering and other machine learning techniques. We reinterpret “contrast-loss” as a blessing. Re-deriving “contrast-loss” using the law of large numbers, we show it results in a distribution’s instances concentrating on a thin “hyper-shell”. The hollow center means apparently chaotically overlapping distributions are actually intrinsically separable. We use this to develop distribution-clustering, an elegant algorithm for grouping of data points by their (unknown) underlying distribution. Distribution-clustering, creates notably clean clusters from raw unlabeled data, estimates the number of clusters for itself and is inherently robust to “outliers” which form their own clusters. This enables trawling for patterns in unorganized data and may be the key to enabling machine intelligence.

****

Eliminating Background-Bias for Robust Person Re-Identification

Maoqing Tian, Shuai Yi, Hongsheng Li, Shihua Li, Xuesen Zhang, Jianping Shi, Junjie Yan, Xiaogang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5794-5803

Person re-identification is an important topic in intelligent surveillance and computer vision. It aims to accurately measure visual similarities between person images for determining whether two images correspond to the same person. State-of-the-art methods mainly utilize deep learning based approaches for learning visual features for describing person appearances. However, we observe that existing deep learning models are biased to capture too much relevance between background appearances of person images. We design a series of experiments with newly created datasets to validate the influence of background information. To solve the background bias problem, we propose a person-region guided pooling deep neural network based on human parsing maps to learn more discriminative person-part features, and propose to augment training data with person images with random background. Extensive experiments demonstrate the robustness and effectiveness of our proposed method.

****

Learning to Evaluate Image Captioning

Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, Serge Belongie; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5804-5812

Evaluation metrics for image captioning face two challenges. Firstly, commonly used metrics such as CIDEr, METEOR, ROUGE and BLEU often do not correlate well with human judgments. Secondly, each metric has well known blind spots to pathological caption constructions, and rule-based metrics lack provisions to repair such blind spots once identified. For example, the newly proposed SPICE correlates well with human judgments, but fails to capture the syntactic structure of a sentence. To address these two challenges, we propose a novel learning based discriminative evaluation metric that is directly trained to distinguish between human and machine-generated captions. In addition, we further propose a data augmentation scheme to explicitly incorporate pathological transformations as negative examples during training. The proposed metric is evaluated with three kinds of robustness tests and its correlation with human judgments. Extensive experiments show that the proposed data augmentation scheme not only makes our metric more robust toward several pathological transformations, but also improves its correlation with human judgments. Our metric outperforms other metrics on both caption level human correlation in Flickr 8k and system level human correlation in COCO. The proposed approach could be served as a learning based evaluation metric that is complementary to existing rule-based metrics.

****

Single-Shot Object Detection With Enriched Semantics

Zhishuai Zhang, Siyuan Qiao, Cihang Xie, Wei Shen, Bo Wang, Alan L. Yuille; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5813-5821

We propose a novel single shot object detection network named Detection with Enriched Semantics (DES). Our motivation is to enrich the semantics of object detection features within a typical deep detector, by a semantic segmentation branch and a global activation module. The segmentation branch is supervised by weak segmentation ground-truth, i.e., no extra annotation is required. In conjunction with that, we employ a global activation module which learns relationship between channels and object classes in a self-supervised manner. Comprehensive experimental results on both PASCAL VOC and MS COCO detection datasets demonstrate the effectiveness of the proposed method. In particular, with a VGG16 based DES, we achieve an mAP of 81.7 on VOC2007 test and an mAP of 32.8 on COCO test-dev with an inference speed of 31.5 milliseconds per image on a Titan Xp GPU. With a lower resolution version, we achieve an mAP of 79.7 on VOC2007 with an inference speed of 13.0 milliseconds per image.

****

Low-Shot Learning With Imprinted Weights

Hang Qi, Matthew Brown, David G. Lowe; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5822-5830

Human vision is able to immediately recognize novel visual categories after seeing just one or a few training examples. We describe how to add a similar capability to ConvNet classifiers by directly setting the final layer weights from novel training examples during low-shot learning. We call this process weight imprinting as it directly sets weights for a new category based on an appropriately scaled copy of the embedding layer activations for that training example. The imprinting process provides a valuable complement to training with stochastic gradient descent, as it provides immediate good classification performance and an initialization for any further fine-tuning in the future. We show how this imprinting process is related to proxy-based embeddings. However, it differs in that only a single imprinted weight vector is learned for each novel category, rather than relying on a nearest-neighbor distance to training instances as typically used with embedding methods. Our experiments show that using averaging of imprinted weights provides better generalization than using nearest-neighbor instance embeddings.

****

Neural Motifs: Scene Graph Parsing With Global Context

Rowan Zellers, Mark Yatskar, Sam Thomson, Yejin Choi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5831-5840

We investigate the problem of producing structured graph representations of visual scenes. Our work analyzes the role of motifs: regularly appearing substructures in scene graphs. We present new quantitative insights on such repeated structures in the Visual Genome dataset. Our analysis shows that object labels are highly predictive of relation labels but not vice-versa. We also find that there are recurring patterns even in larger subgraphs: more than 50% of graphs contain motifs involving at least two relations. Our analysis motivates a new baseline: given object detections, predict the most frequent relation between object pairs with the given labels, as seen in the training set. This baseline improves on the previous state-of-the-art by an average of 3.6% relative improvement across evaluation settings. We then introduce Stacked Motif Networks, a new architecture designed to capture higher order motifs in scene graphs that further improves over our strong baseline by an average 7.1% relative gain. Our code is available at github.com/rowanz/neural-motifs.

****

Variational Autoencoders for Deforming 3D Mesh Models

Qingyang Tan, Lin Gao, Yu-Kun Lai, Shihong Xia; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5841-5850

3D geometric contents are becoming increasingly popular. In this paper, we study the problem of analyzing deforming 3D meshes using deep neural networks. Deforming 3D meshes are flexible to represent 3D animation sequences as well as collections of objects of the same category, allowing diverse shapes with large-scale non-linear deformations. We propose a novel framework which we call mesh variational autoencoders (mesh VAE), to explore the probabilistic latent space of 3D surfaces. The framework is easy to train, and requires very few training examples. We also propose an extended model which allows flexibly adjusting the significance of different latent variables by altering the prior distribution. Extensive experiments demonstrate that our general framework is able to learn a reasonable representation for a collection of deformable shapes, and produce competitive results for a variety of applications, including shape generation, shape interpolation, shape space embedding and shape exploration, outperforming state-of-the-art methods.

****

Fast Monte-Carlo Localization on Aerial Vehicles Using Approximate Continuous Belief Representations

Aditya Dhawale, Kumar Shaurya Shankar, Nathan Michael; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5851-5859

Size, weight, and power constrained platforms impose constraints on computationa l resources that introduce unique challenges in implementing localization algorithms. We present a framework to perform fast localization on such platforms enabled by the compressive capabilities of Gaussian Mixture Model representations of point cloud data. Given raw structural data from a depth sensor and pitch and roll estimates from an on-board attitude reference system, a multi-hypothesis particle filter localizes the vehicle by exploiting the likelihood of the data originating from the mixture model. We demonstrate analysis of this likelihood in the vicinity of the ground truth pose and detail its utilization in a particle filter-based vehicle localization strategy, and later present results of real-time implementations on a desktop system and an off-the-shelf embedded platform that outperform localization results from running a state-of-the-art algorithm on the same environment.

****

DeLS-3D: Deep Localization and Segmentation With a 3D Semantic Map

Peng Wang, Ruigang Yang, Binbin Cao, Wei Xu, Yuanqing Lin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5860-5869

For applications such as augmented reality, autonomous driving, self-localization/camera pose estimation and scene parsing are crucial technologies. In this paper, we propose a unified framework to tackle these two problems simultaneously. The uniqueness of our design is a sensor fusion scheme which integrates camera videos, motion sensors (GPS/IMU), and a 3D semantic map in order to achieve robustness and efficiency of the system.Specifically, we first have an initial coarse camera pose obtained from consumer-grade GPS/IMU, based on which a label map can be rendered from the 3D semantic map. Then, the rendered label map and the RGB image are jointly fed into a pose CNN, yielding a corrected camera pose. In addition, to incorporate temporal information, a multi-layer recurrent neural network (RNN) is further deployed improve the pose accuracy. Finally, based on the pose from RNN, we render a new label map, which is fed together with the RGB image into a segment CNN which produces per-pixel semantic label. In order to validate our approach, we build a dataset with registered 3D point clouds and video camera images. Both the point clouds and the images are semantically-labeled. Each video frame has ground truth pose from highly accurate motion sensors. We show that practically, pose estimation solely relying on images like PoseNet~cite{Kendall_2015_ICCV} may fail due to street view confusion, and it is important to fuse multiple sensors. Finally, various ablation studies are performed, which demonstrate the effectiveness of the proposed system. In particular, we show that scene parsing and pose estimation are mutually beneficial to achieve a more robust and accurate system.

****

LiDAR-Video Driving Dataset: Learning Driving Policies Effectively

Yiping Chen, Jingkang Wang, Jonathan Li, Cewu Lu, Zhipeng Luo, Han Xue, Cheng Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5870-5878

Learning autonomous-driving policies is one of the most challenging but promising tasks for computer vision. Most researchers believe that future research and applications should combine cameras, video recorders and laser scanners to obtain comprehensive semantic understanding of real traffic. However, current approaches only learn from large-scale videos, due to the lack of benchmarks that consist of precise laser-scanner data. In this paper, we are the first to propose a LiDAR-Video dataset, which provides large-scale high-quality point clouds scanned by a Velodyne laser, videos recorded by a dashboard camera and standard drivers' behaviors. Extensive experiments demonstrate that extra depth information help networks to determine driving policies indeed.

****

Logo Synthesis and Manipulation With Clustered Generative Adversarial Networks

Alexander Sage, Eirikur Agustsson, Radu Timofte, Luc Van Gool; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5879-5888

Designing a logo for a new brand is a lengthy and tedious back-and-forth process between a designer and a client. In this paper we explore to what extent machine learning can solve the creative task of the designer. For this, we build a dataset -- LLD -- of 600k+ logos crawled from the world wide web. Training Generative Adversarial Networks (GANs) for logo synthesis on such multi-modal data is not straightforward and results in mode collapse for some state-of-the-art methods. We propose the use of synthetic labels obtained through clustering to disentangle and stabilize GAN training, and validate this approach on CIFAR-10 and ImageNet-small to demonstrate its generality. We are able to generate a high diversity of plausible logos and demonstrate latent space exploration techniques to ease the logo design task in an interactive manner. GANs can cope with multi-modal data by means of synthetic labels achieved through clustering, and our results show the creative potential of such techniques for logo synthesis and manipulation. Our dataset and models are publicly available at https://data.vision.ee.ethz.ch/sagea/lld.

****

Egocentric Basketball Motion Planning From a Single First-Person Image

Gedas Bertasius, Aaron Chan, Jianbo Shi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5889-5898

We present a model that uses a single first-person image to generate an egocentric basketball motion sequence in the form of a 12D camera configuration trajectory, which encodes a player's 3D location and 3D head orientation throughout the sequence. To do this, we first introduce a future convolutional neural network (CNN) that predicts an initial sequence of 12D camera configurations, aiming to capture how real players move during a one-on-one basketball game. We also introduce a goal verifier network, which is trained to verify that a given camera configuration is consistent with the final goals of real one-on-one basketball players. Next, we propose an inverse synthesis procedure to synthesize a refined sequence of 12D camera configurations that (1) sufficiently matches the initial configurations predicted by the future CNN, while (2) maximizing the output of the goal verifier network. Finally, by following the trajectory resulting from the refined camera configuration sequence, we obtain the complete 12D motion sequence. Our model generates realistic basketball motion sequences that capture the goals of real players, outperforming standard deep learning approaches such as recurrent neural networks (RNNs), long short-term memory networks (LSTMs), and generative adversarial networks (GANs).

****

Human-Centric Indoor Scene Synthesis Using Stochastic Grammar

Siyuan Qi, Yixin Zhu, Siyuan Huang, Chenfanfu Jiang, Song-Chun Zhu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5899-5908

We present a human-centric method to sample and synthesize 3D room layouts and 2D images thereof, for the purpose of obtaining large-scale 2D/3D image data with the perfect per-pixel ground truth. An attributed spatial And-Or graph (S-AOG) is proposed to represent indoor scenes. The S-AOG is a probabilistic grammar model, in which the terminal nodes are object entities including room, furniture, and supported objects. Human contexts as contextual relations are encoded by Markov Random Fields (MRF) on the terminal nodes. We learn the distributions from an indoor scene dataset and sample new layouts using Monte Carlo Markov Chain. Experiments demonstrate that the proposed method can robustly sample a large variety of realistic room layouts based on three criteria: (i) visual realism comparing to a state-of-the-art room arrangement method, (ii) accuracy of the affordance maps with respect to ground-truth, and (ii) the functionality and naturalness of synthesized rooms evaluated by human subjects.

****

Rotation-Sensitive Regression for Oriented Scene Text Detection

Minghui Liao, Zhen Zhu, Baoguang Shi, Gui-song Xia, Xiang Bai; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5909-5918

Text in natural images is of arbitrary orientations, requiring detection in terms of oriented bounding boxes. Normally, a multi-oriented text detector often inv olves two key tasks: 1) text presence detection, which is a classification problem disregarding text orientation; 2) oriented bounding box regression, which concerns about text orientation. Previous methods rely on shared features for both tasks, resulting in degraded performance due to the incompatibility of the two tasks. To address this issue, we propose to perform classification and regression on features of different characteristics, extracted by two network branches of different designs. Concretely, the regression branch extracts rotation-sensitive features by actively rotating the convolutional filters, while the classification branch extracts rotation-invariant features by pooling the rotation-sensitive features. The proposed method named Rotation-sensitive Regression Detector (RRD) achieves state-of-the-art performance on several oriented scene text benchmark datasets, including ICDAR 2015, MSRA-TD500, RCTW-17, and COCO-Text. Furthermore, RRD achieves a significant improvement on a ship collection dataset, demonstrating its generality on oriented object detection.

****

Separating Self-Expression and Visual Content in Hashtag Supervision

Andreas Veit, Maximilian Nickel, Serge Belongie, Laurens van der Maaten; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5919-5927

The variety, abundance, and structured nature of hashtags make them an interesting data source for training vision models. For instance, hashtags have the potential to significantly reduce the problem of manual supervision and annotation when learning vision models for a large number of concepts. However, a key challenge when learning from hashtags is that they are inherently subjective because they are provided by users as a form of self-expression. As a consequence, hashtags may have synonyms (different hashtags referring to the same visual content) and may be polysemous (the same hashtag referring to different visual content). These challenges limit the effectiveness of approaches that simply treat hashtags as image-label pairs. This paper presents an approach that extends upon modeling simple image-label pairs with a joint model of images, hashtags, and users. We demonstrate the efficacy of such approaches in image tagging and retrieval experiments, and show how the joint model can be used to perform user-conditional retrieval and tagging.

****

Distort-and-Recover: Color Enhancement Using Deep Reinforcement Learning

Jongchan Park, Joon-Young Lee, Donggeun Yoo, In So Kweon; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5928-5936

Learning-based color enhancement approaches typically learn to map from input images to retouched images. Most of existing methods require expensive pairs of input-retouched images or produce results in a non-interpretable way. In this paper, we present a deep reinforcement learning (DRL) based method for color enhancement to explicitly model the step-wise nature of human retouching process. We cast a color enhancement process as a Markov Decision Process where actions are defined as global color adjustment operations. Then we train our agent to learn the optimal global enhancement sequence of the actions. In addition, we present a `distort-and-recover' training scheme which only requires high-quality reference images for training instead of input and retouched image pairs. Given high-quality reference images, we distort the images' color distribution and form distorted-reference image pairs for training. Through extensive experiments, we show that our method produces decent enhancement results and our DRL approach is more suitable for the `distort-and-recover' training scheme than previous supervised approaches. Supplementary material and code are available at https://sites.google.com/view/distort-and-recover/

****

Im2Flow: Motion Hallucination From Static Images for Action Recognition

Ruohan Gao, Bo Xiong, Kristen Grauman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5937-5947

Existing methods to recognize actions in static images take the images at their face value, learning the appearances---objects, scenes, and body poses---that di

stinguish each action class. However, such models are deprived of the rich dynamic structure and motions that also define human activity. We propose an approach that hallucinates the unobserved future motion implied by a single snapshot to help static-image action recognition. The key idea is to learn a prior over short-term dynamics from thousands of unlabeled videos, infer the anticipated optical flow on novel static images, and then train discriminative models that exploit both streams of information. Our main contributions are twofold. First, we devise an encoder-decoder convolutional neural network and a novel optical flow encoding that can translate a static image into an accurate flow map. Second, we show the power of hallucinated flow for recognition, successfully transferring the learned motion into a standard two-stream network for activity recognition. On seven datasets, we demonstrate the power of the approach. It not only achieves state-of-the-art accuracy for dense optical flow prediction, but also consistently enhances recognition of actions and dynamic scenes.
****
Finding "It": Weakly-Supervised Reference-Aware Visual Grounding in Instructional Videos
De-An Huang, Shyamal Buch, Lucio Dery, Animesh Garg, Li Fei-Fei, Juan Carlos Niebles; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5948-5957
Grounding textual phrases in visual content with standalone image-sentence pairs is a challenging task. When we consider grounding in instructional videos, this problem becomes profoundly more complex: the latent temporal structure of instructional videos breaks independence assumptions and necessitates contextual understanding for resolving ambiguous visual-linguistic cues. Furthermore, dense annotations and video data scale mean supervised approaches are prohibitively costly. In this work, we propose to tackle this new task with a weakly-supervised framework for reference-aware visual grounding in instructional videos, where only the temporal alignment between the transcription and the video segment are available for supervision. We introduce the visually grounded action graph, a structured representation capturing the latent dependency between grounding and references in video. For optimization, we propose a new reference-aware multiple instance learning (RA-MIL) objective for weak supervision of grounding in videos. We evaluate our approach over unconstrained videos from YouCookII and RoboWatch, augmented with new reference-grounding test set annotations. We demonstrate that our jointly optimized, reference-aware approach simultaneously improves visual grounding, reference-resolution, and generalization to unseen instructional video categories.
****
Actor and Action Video Segmentation From a Sentence
Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, Cees G. M. Snoek; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5958-5966
This paper strives for pixel-level segmentation of actors and their actions in video content. Different from existing works, which all learn to segment from a fixed vocabulary of actor and action pairs, we infer the segmentation from a natural language input sentence. This allows to distinguish between fine-grained actors in the same super-category, identify actor and action instances, and segment pairs that are outside of the actor and action vocabulary. We propose a fully-convolutional model for pixel-level actor and action segmentation using an encoder-decoder architecture optimized for video. To show the potential of actor and action video segmentation from a sentence, we extend two popular actor and action datasets with more than 7,500 natural language descriptions. Experiments demonstrate the quality of the sentence-guided segmentations, the generalization ability of our model, and its advantage for traditional actor and action segmentation compared to the state-of-the-art.
****
Egocentric Activity Recognition on a Budget
Rafael Possas, Sheila Pinto Caceres, Fabio Ramos; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5967-5976

Recent advances in embedded technology have enabled more pervasive machine learning. One of the common applications in this field is Egocentric Activity Recognition (EAR), where users wearing a device such as a smartphone or smartglasses are able to receive feedback from the embedded device. Recent research on activity recognition has mainly focused on improving accuracy by using resource intensive techniques such as multi-stream deep networks. Although this approach has provided state-of-the-art results, in most cases it neglects the natural resource constraints (e.g. battery) of wearable devices. We develop a Reinforcement Learning model-free method to learn energy-aware policies that maximize the use of low-energy cost predictors while keeping competitive accuracy levels. Our results show that a policy trained on an egocentric dataset is able use the synergy between motion sensors and vision to effectively tradeoff energy expenditure and accuracy on smartglasses operating in realistic, real-world conditions.

****

CNN in MRF: Video Object Segmentation via Inference in a CNN-Based Higher-Order Spatio-Temporal MRF

Linchao Bao, Baoyuan Wu, Wei Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5977-5986

This paper addresses the problem of video object segmentation, where the initial object mask is given in the first frame of an input video. We propose a novel spatio-temporal Markov Random Field (MRF) model defined over pixels to handle this problem. Unlike conventional MRF models, the spatial dependencies among pixels in our model are encoded by a Convolutional Neural Network (CNN). Specifically, for a given object, the probability of a labeling to a set of spatially neighboring pixels can be predicted by a CNN trained for this specific object. As a result, higher-order, richer dependencies among pixels in the set can be implicitly modeled by the CNN. With temporal dependencies established by optical flow, the resulting MRF model combines both spatial and temporal cues for tackling video object segmentation. However, performing inference in the MRF model is very difficult due to the very high-order dependencies. To this end, we propose a novel CNN-embedded algorithm to perform approximate inference in the MRF. This algorithm proceeds by alternating between a temporal fusion step and a feed-forward CNN step. When initialized with an appearance-based one-shot segmentation CNN, our model outperforms the winning entries of the DAVIS 2017 Challenge, without resorting to model ensembling or any dedicated detectors.

****

Action Sets: Weakly Supervised Action Segmentation Without Ordering Constraints

Alexander Richard, Hilde Kuehne, Juergen Gall; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5987-5996

Action detection and temporal segmentation of actions in videos are topics of increasing interest. While fully supervised systems have gained much attention lately, full annotation of each action within the video is costly and impractical for large amounts of video data. Thus, weakly supervised action detection and temporal segmentation methods are of great importance. While most works in this area assume an ordered sequence of occurring actions to be given, our approach only uses a set of actions. Such action sets provide much less supervision since neither action ordering nor the number of action occurrences are known. In exchange, they can be easily obtained, for instance, from meta-tags, while ordered sequences still require human annotation. We introduce a system that automatically learns to temporally segment and label actions in a video, where the only supervision that is used are action sets. An evaluation on three datasets shows that our method still achieves good results although the amount of supervision is significantly smaller than for other related methods.

****

Low-Latency Video Semantic Segmentation

Yule Li, Jianping Shi, Dahua Lin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5997-6005

Recent years have seen remarkable progress in semantic segmentation. Yet, it remains a challenging task to apply segmentation techniques to video-based applications. Specifically, the high throughput of video streams, the sheer cost of runn ing fully convolutional networks, together with the low-latency requirements in many real-world applications, e.g. autonomous driving, present a significant challenge to the design of the video segmentation framework. To tackle this combined challenge, we develop a framework for video semantic segmentation, which incorporates two novel components:(1) a feature propagation module that adaptively fuses features over time via spatially variant convolution, thus reducing the cost of per-frame computation; and (2) an adaptive scheduler that dynamically allocate computation based on accuracy prediction. Both components work together to ensure low latency while maintaining high segmentation quality. On both Cityscapes and CamVid, the proposed framework obtained competitive performance compared to the state of the art, while substantially reducing the latency, from 360 ms to 119 ms.

****

Fine-Grained Video Captioning for Sports Narrative

Huanyu Yu, Shuo Cheng, Bingbing Ni, Minsi Wang, Jian Zhang, Xiaokang Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6006-6015

Despite recent emergence of video caption methods, how to generate fine-grained video descriptions (i.e., long and detailed commentary about individual movements of multiple subjects as well as their frequent interactions) is far from being solved, which however has great applications such as automatic sports narrative. To this end, this work makes the following contributions. First, to facilitate this novel research of fine-grained video caption, we collected a novel dataset called Fine-grained Sports Narrative dataset (FSN) that contains 2K sports videos with ground-truth narratives from YouTube.com. Second, we develop a novel performance evaluation metric named Fine-grained Captioning Evaluation (FCE) to cope with this novel task. Considered as an extension of the widely used METEOR, it measures not only the linguistic performance but also whether the action details and their temporal orders are correctly described. Third, we propose a new framework for fine-grained sports narrative task. This network features three branches: 1) a spatio-temporal entity localization and role discovering sub-network; 2) a fine-grained action modeling sub-network for local skeleton motion description; and 3) a group relationship modeling sub-network to model interactions between players. We further fuse the features and decode them into long narratives by a hierarchically recurrent structure. Extensive experiments on the FSN dataset demonstrates the validity of the proposed framework for fine-grained video caption.

****

End-to-End Learning of Motion Representation for Video Understanding

Lijie Fan, Wenbing Huang, Chuang Gan, Stefano Ermon, Boqing Gong, Junzhou Huang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6016-6025

Despite the recent success of end-to-end learned representations, hand-crafted optical flow features are still widely used in video analysis tasks. To fill this gap, we propose TVNet, a novel end-to-end trainable neural network, to learn optical-flow-like features from data. TVNet subsumes a specific optical flow solver, the TV-L1 method, and is initialized by unfolding its optimization iterations as neural layers. TVNet can therefore be used directly without any extra learning. Moreover, it can be naturally concatenated with other task-specific networks to formulate an end-to-end architecture, thus making our method more efficient than current multi-stage approaches by avoiding the need to pre-compute and store features on disk. Finally, the parameters of the TVNet can be further fine-tuned by end-to-end training. This enables TVNet to learn richer and task-specific patterns beyond exact optical flow. Extensive experiments on two action recognition benchmarks verify the effectiveness of the proposed approach. Our TVNet achieves better accuracies than all compared methods, while being competitive with the fastest counterpart in terms of features extraction time.

****

Compressed Video Action Recognition

Chao-Yuan Wu, Manzil Zaheer, Hexiang Hu, R. Manmatha, Alexander J. Smola, Philip p Krähenbühl; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6026-6035

Training robust deep video representations has proven to be much more challenging than learning deep image representations. This is in part due to the enormous size of raw video streams and the high temporal redundancy; the true and interesting signal is often drowned in too much irrelevant data. Motivated by that the superfluous information can be reduced by up to two orders of magnitude by video compression (using H.264, HEVC, etc.), we propose to train a deep network directly on the compressed video. This representation has a higher information density, and we found the training to be easier. In addition, the signals in a compressed video provide free, albeit noisy, motion information. We propose novel techniques to use them effectively. Our approach is about 4.6 times faster than Res3D and 2.7 times faster than ResNet-152. On the task of action recognition, our approach outperforms all the other methods on the UCF-101, HMDB-51, and Charades dataset.

****

Features for Multi-Target Multi-Camera Tracking and Re-Identification

Ergys Ristani, Carlo Tomasi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6036-6046

Multi-Target Multi-Camera Tracking (MTMCT) tracks many people through video taken from several cameras. Person Re-Identification (Re-ID) retrieves from a gallery images of people similar to a person query image. We learn good features for both MTMCT and Re-ID with a convolutional neural network. Our contributions include an adaptive weighted triplet loss for training and a new technique for hard-identity mining. Our method outperforms the state of the art both on the DukeMTMC benchmarks for tracking, and on the Market-1501 and DukeMTMC-ReID benchmarks for Re-ID. We examine the correlation between good Re-ID and good MTMCT scores, and perform ablation studies to elucidate the contributions of the main components of our system. Code is available.

****

AVA: A Video Dataset of Spatio-Temporally Localized Atomic Visual Actions

Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, Jitendra Malik; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6047-6056

This paper introduces a video dataset of spatio-temporally localized Atomic Visual Actions (AVA). The AVA dataset densely annotates 80 atomic visual actions in 437 15-minute video clips, where actions are localized in space and time, resulting in 1.59M action labels with multiple labels per person occurring frequently. The key characteristics of our dataset are: (1) the definition of atomic visual actions, rather than composite actions; (2) precise spatio-temporal annotations with possibly multiple annotations for each person; (3) exhaustive annotation of these atomic actions over 15-minute video clips; (4) people temporally linked across consecutive segments; and (5) using movies to gather a varied set of action representations. This departs from existing datasets for spatio-temporal action recognition, which typically provide sparse annotations for composite actions in short video clips. AVA, with its realistic scene and action complexity, exposes the intrinsic difficulty of action recognition. To benchmark this, we present a novel approach for action localization that builds upon the current state-of-the-art methods, and demonstrates better performance on JHMDB and UCF101-24 categories. While setting a new state of the art on existing datasets, the overall results on AVA are low at 15.8% mAP, underscoring the need for developing new approaches for video understanding.

****

Who's Better? Who's Best? Pairwise Deep Ranking for Skill Determination

Hazel Doughty, Dima Damen, Walterio Mayol-Cuevas; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6057-6066

This paper presents a method for assessing skill from video, applicable to a variety of tasks, ranging from surgery to drawing and rolling pizza dough. We formulate the problem as pairwise (who’s better?) and overall (who’s best?) ranking o f video collections, using supervised deep ranking. We propose a novel loss function that learns discriminative features when a pair of videos exhibit variance in skill, and learns shared features when a pair of videos exhibit comparable skill levels. Results demonstrate our method is applicable across tasks, with the percentage of correctly ordered pairs of videos ranging from 70% to 83% for four datasets. We demonstrate the robustness of our approach via sensitivity analysis of its parameters. We see this work as effort toward the automated organization of how-to video collections and overall, generic skill determination in video.

****

MX-LSTM: Mixing Tracklets and Vislets to Jointly Forecast Trajectories and Head Poses

Irtiza Hasan, Francesco Setti, Theodore Tsesmelis, Alessio Del Bue, Fabio Galasso, Marco Cristani; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6067-6076

Recent approaches on trajectory forecasting use tracklets to predict the future positions of pedestrians exploiting Long Short Term Memory (LSTM) architectures. This paper shows that adding vislets, that is, short sequences of head pose estimations, allows to increase significantly the trajectory forecasting performance. We then propose to use vislets in a novel framework called MX-LSTM, capturing the interplay between tracklets and vislets thanks to a joint unconstrained optimization of full covariance matrices during the LSTM backpropagation. At the same time, MX-LSTM predicts the future head poses, increasing the standard capabilities of the long-term trajectory forecasting approaches. With standard head pose estimators and an attentional-based social pooling, Mixing-LSTM scores the new trajectory forecasting state-of-the-art in all the considered datasets (Zara01, Zara02, UCY, and TownCentre) with a dramatic margin when the pedestrians slow down, a case where most of the forecasting approaches struggle to provide an accurate solution.

****

Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, Lei Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6077-6086

Top-down visual attention mechanisms have been used extensively in image captioning and visual question answering (VQA) to enable deeper image understanding through fine-grained analysis and even multiple steps of reasoning. In this work, we propose a combined bottom-up and top-down attention mechanism that enables attention to be calculated at the level of objects and other salient image regions. This is the natural basis for attention to be considered. Within our approach, the bottom-up mechanism (based on Faster R-CNN) proposes image regions, each with an associated feature vector, while the top-down mechanism determines feature weightings. Applying this approach to image captioning, our results on the MSCOCO test server establish a new state-of-the-art for the task, achieving CIDEr / SPICE / BLEU-4 scores of 117.9, 21.5 and 36.9, respectively. Demonstrating the broad applicability of the method, applying the same approach to VQA we obtain first place in the 2017 VQA Challenge.

****

Improved Fusion of Visual and Language Representations by Dense Symmetric Co-Attention for Visual Question Answering

Duy-Kien Nguyen, Takayuki Okatani; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6087-6096

A key solution to visual question answering (VQA) exists in how to fuse visual and language features extracted from an input image and question. We show that an attention mechanism that enables dense, bi-directional interactions between the two modalities contributes to boost accuracy of prediction of answers. Specifically, we present a simple architecture that is fully symmetric between visual and language representations, in which each question word attends on image regions and each image region attends on question words. It can be stacked to form a hierarchy for multi-step interactions between an image-question pair. We show thro ugh experiments that the proposed architecture achieves a new state-of-the-art on VQA and VQA 2.0 despite its small size. We also present qualitative evaluation, demonstrating how the proposed attention mechanism can generate reasonable attention maps on images and questions, which leads to the correct answer prediction.

****

FlipDial: A Generative Model for Two-Way Visual Dialogue

Daniela Massiceti, N. Siddharth, Puneet K. Dokania, Philip H.S. Torr; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6097-6105

We present FlipDial, a generative model for Visual Dialogue that simultaneously plays the role of both participants in a visually-grounded dialogue. Given context in the form of an image and an associated caption summarising the contents of the image, FlipDial learns both to answer questions and put forward questions, capable of generating entire sequences of dialogue (question-answer pairs) which are diverse and relevant to the image. To do this, FlipDial relies on a simple but surprisingly powerful idea: it uses convolutional neural networks (CNNs) to encode entire dialogues directly, implicitly capturing dialogue context, and conditional VAEs to learn the generative model. FlipDial outperforms the state-of-the-art model in the sequential answering task (1VD) on the VisDial dataset by 5 points in Mean Rank using the generated answers. We are the first to extend this paradigm to full two-way visual dialogue (2VD), where our model is capable of generating both questions and answers in sequence based on a visual input, for which we propose a set of novel evaluation measures and metrics.

****

Are You Talking to Me? Reasoned Visual Dialog Generation Through Adversarial Learning

Qi Wu, Peng Wang, Chunhua Shen, Ian Reid, Anton van den Hengel; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6106-6115

The Visual Dialogue task requires an agent to engage in a conversation about an image with a human. It represents an extension of the Visual Question Answering task in that the agent needs to answer a question about an image, but it needs to do so in light of the previous dialogue that has taken place. The key challenge in Visual Dialogue is thus maintaining a consistent, and natural dialogue while continuing to answer questions correctly. We present a novel approach that combines Reinforcement Learning and Generative Adversarial Networks (GANs) to generate more human-like responses to questions. The GAN helps overcome the relative paucity of training data, and the tendency of the typical MLE-based approach to generate overly terse answers. Critically, the GAN is tightly integrated into the attention mechanism that generates human-interpretable reasons for each answer. This means that the discriminative model of the GAN has the task of assessing whether a candidate answer is generated by a human or not, given the provided reason. This is significant because it drives the generative model to produce high quality answers that are well supported by the associated reasoning. The method also generates the state-of-the-art results on the primary benchmark.

****

Visual Question Generation as Dual Task of Visual Question Answering

Yikang Li, Nan Duan, Bolei Zhou, Xiao Chu, Wanli Ouyang, Xiaogang Wang, Ming Zhou; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6116-6124

Visual question answering (VQA) and visual question generation (VQG) are two trending topics in the computer vision, but they are usually explored separately despite their intrinsic complementary relationship. In this paper, we propose an end-to-end unified model, the Invertible Question Answering Network (iQAN), to introduce question generation as a dual task of question answering to improve the VQA performance. With our proposed invertible bilinear fusion module and parameter sharing scheme, our iQAN can accomplish VQA and its dual task VQG simultaneously. By jointly trained on two tasks with our proposed dual regularizers~(termed as Dual Training), our model has a better understanding of the interactions amo ng images, questions and answers. After training, iQAN can take either question or answer as input, and output the counterpart. Evaluated on the CLEVR and VQA2 datasets, our iQAN improves the top-1 accuracy of the prior art MUTAN VQA method by 1.33% and 0.88% (absolute increase). We also show that our proposed dual training framework can consistently improve model performances of many popular VQA architectures.

****

Unsupervised Textual Grounding: Linking Words to Image Concepts

Raymond A. Yeh, Minh N. Do, Alexander G. Schwing; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6125-6134

Textual grounding, i.e., linking words to objects in images, is a challenging but important task for robotics and human-computer interaction. Existing techniques benefit from recent progress in deep learning and generally formulate the task as a supervised learning problem, selecting a bounding box from a set of possible options. To train these deep net based approaches, access to a large-scale datasets is required, however, constructing such a dataset is time-consuming and expensive. Therefore, we develop a completely unsupervised mechanism for textual grounding using hypothesis testing as a mechanism to link words to detected image concepts. We demonstrate our approach on the ReferIt Game dataset and the Flickr30k data, outperforming baselines by 7.98% and 6.96% respectively.

****

Focal Visual-Text Attention for Visual Question Answering

Junwei Liang, Lu Jiang, Liangliang Cao, Li-Jia Li, Alexander G. Hauptmann; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6135-6143

Recent insights on language and vision with neural networks have been successfully applied to simple single-image visual question answering. However, to tackle real-life question answering problems on multimedia collections such as personal photos, we have to look at whole collections with sequences of photos or videos. When answering questions from a large collection, a natural problem is to identify snippets to support the answer. In this paper, we describe a novel neural network called Focal Visual-Text Attention network (FVTA) for collective reasoning in visual question answering, where both visual and text sequence information such as images and text metadata are presented. FVTA introduces an end-to-end approach that makes use of a hierarchical process to dynamically determine what media and what time to focus on in the sequential data to answer the question. FVTA can not only answer the questions well but also provides the justifications which the system results are based upon to get the answers. FVTA achieves state-of-the-art performance on the MemexQA dataset and competitive results on the MovieQA dataset.

****

SeGAN: Segmenting and Generating the Invisible

Kiana Ehsani, Roozbeh Mottaghi, Ali Farhadi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6144-6153

Objects often occlude each other in scenes; Inferring their appearance beyond their visible parts plays an important role in scene understanding, depth estimation, object interaction and manipulation. In this paper, we study the challenging problem of completing the appearance of occluded objects. Doing so requires knowing which pixels to paint (segmenting the invisible parts of objects) and what color to paint them (generating the invisible parts). Our proposed novel solution, SeGAN, jointly optimizes for both segmentation and generation of the invisible parts of objects. Our experimental results show that: (a) SeGAN can learn to generate the appearance of the occluded parts of objects; (b) SeGAN outperforms state-of-the-art segmentation baselines for the invisible parts of objects; (c) trained on synthetic photo realistic images, SeGAN can reliably segment natural images; (d) by reasoning about occluder-occludee relations, our method can infer depth layering.

****

Cascade R-CNN: Delving Into High Quality Object Detection

Zhaowei Cai, Nuno Vasconcelos; Proceedings of the IEEE Conference on Computer Vi sion and Pattern Recognition (CVPR), 2018, pp. 6154-6162

In object detection, an intersection over union (IoU) threshold is required to define positives and negatives. An object detector, trained with low IoU threshold, e.g. 0.5, usually produces noisy detections. However, detection performance tends to degrade with increasing the IoU thresholds. Two main factors are responsible for this: 1) overfitting during training, due to exponentially vanishing positive samples, and 2) inference-time mismatch between the IoUs for which the detector is optimal and those of the input hypotheses. A multi-stage object detection architecture, the Cascade R-CNN, is proposed to address these problems. It consists of a sequence of detectors trained with increasing IoU thresholds, to be sequentially more selective against close false positives. The detectors are trained stage by stage, leveraging the observation that the output of a detector is a good distribution for training the next higher quality detector. The resampling of progressively improved hypotheses guarantees that all detectors have a positive set of examples of equivalent size, reducing the overfitting problem. The same cascade procedure is applied at inference, enabling a closer match between the hypotheses and the detector quality of each stage. A simple implementation of the Cascade R-CNN is shown to surpass all single-model object detectors on the challenging COCO dataset. Experiments also show that the Cascade R-CNN is widely applicable across detector architectures, achieving consistent gains independently of the baseline detector strength. The code is available at https://github.com/zhaoweicai/cascade-rcnn.

****

Learning Semantic Concepts and Order for Image and Sentence Matching

Yan Huang, Qi Wu, Chunfeng Song, Liang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6163-6171

Image and sentence matching has made great progress recently, but it remains challenging due to the large visual semantic discrepancy. This mainly arises from that the representation of pixel-level image usually lacks of high-level semantic information as in its matched sentence. In this work, we propose a semantic-enhanced image and sentence matching model, which can improve the image representation by learning semantic concepts and then organizing them in a correct semantic order. Given an image, we first use a multi-regional multi-label CNN to predict its semantic concepts, including objects, properties, actions, etc. Then, considering that different orders of semantic concepts lead to diverse semantic meanings, we use a context-gated sentence generation scheme for semantic order learning. It simultaneously uses the image global context containing concept relations as reference and the groundtruth semantic order in the matched sentence as supervision. After obtaining the improved image representation, we learn the sentence representation with a conventional LSTM, and then jointly perform image and sentence matching and sentence generation for model learning. Extensive experiments demonstrate the effectiveness of our learned semantic concepts and order, by achieving the state-of-the-art results on two public benchmark datasets.

****

Functional Map of the World

Gordon Christie, Neil Fendley, James Wilson, Ryan Mukherjee; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6172-6180

We present a new dataset, Functional Map of the World (fMoW), which aims to inspire the development of machine learning models capable of predicting the functional purpose of buildings and land use from temporal sequences of satellite images and a rich set of metadata features. The metadata provided with each image enables reasoning about location, time, sun angles, physical sizes, and other features when making predictions about objects in the image. Our dataset consists of over 1 million images from over 200 countries. For each image, we provide at least one bounding box annotation containing one of 63 categories, including a "false detection" category. We present an analysis of the dataset along with baseline approaches that reason about metadata and temporal views. Our data, code, and pretrained models have been made publicly available.

****

MegDet: A Large Mini-Batch Object Detector

Chao Peng, Tete Xiao, Zeming Li, Yuning Jiang, Xiangyu Zhang, Kai Jia, Gang Yu, Jian Sun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6181-6189

The development of object detection in the era of deep learning, from R-CNN [11], Fast/Faster R-CNN [10, 31] to recent Mask R-CNN [14] and RetinaNet [24], mainly come from novel network, new framework, or loss design. How- ever, mini-batch size, a key factor for the training of deep neural networks, has not been well studied for object detec- tion. In this paper, we propose a Large Mini-Batch Object Detector (MegDet) to enable the training with a large mini- batch size up to 256, so that we can effectively utilize at most 128 GPUs to significantly shorten the training time. Technically, we suggest a warmup learning rate policy and Cross-GPU Batch Normalization, which together allow us to successfully train a large mini-batch detector in much less time (e.g., from 33 hours to 4 hours), and achieve even better accuracy. The MegDet is the backbone of our sub- mission (mmAP 52.5%) to COCO 2017 Challenge, where we won the 1st place of Detection task.

****

Learning Globally Optimized Object Detector via Policy Gradient

Yongming Rao, Dahua Lin, Jiwen Lu, Jie Zhou; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6190-6198

In this paper, we propose a simple yet effective method to learn globally optimized detector for object detection, which is a simple modification to the standard cross-entropy gradient inspired by the REINFORCE algorithm. In our approach, the cross-entropy gradient is adaptively adjusted according to overall mean Average Precision (mAP) of the current state for each detection candidate, which leads to more effective gradient and global optimization of detection results, and brings no computational overhead. Benefiting from more precise gradients produced by the global optimization method, our framework significantly improves state-of-the-art object detectors. Furthermore, since our method is based on scores and bounding boxes without modification on the architecture of object detector, it can be easily applied to off-the-shelf modern object detection frameworks.

****

Photographic Text-to-Image Synthesis With a Hierarchically-Nested Adversarial Network

Zizhao Zhang, Yuanpu Xie, Lin Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6199-6208

This paper presents a novel method to deal with the challenging task of generating photographic images conditioned on semantic image descriptions. Our method introduces accompanying hierarchical-nested adversarial objectives inside the network hierarchies, which regularize mid-level representations and assist generator training to capture the complex image statistics. We present an extensile single-stream generator architecture to better adapt the jointed discriminators and push generated images up to high resolutions. We adopt a multi-purpose adversarial loss to encourage more effective image and text information usage in order to improve the semantic consistency and image fidelity simultaneously. Furthermore, we introduce a new visual-semantic similarity measure to evaluate the semantic consistency of generated images. With extensive experimental validation on three public datasets, our method significantly improves previous state of the arts on all datasets over different evaluation metrics.

****

Illuminant Spectra-Based Source Separation Using Flash Photography

Zhuo Hui, Kalyan Sunkavalli, Sunil Hadap, Aswin C. Sankaranarayanan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6209-6218

Real-world lighting often consists of multiple illuminants with different spectra. Separating and manipulating these illuminants in post-process is a challenging problem that requires either significant manual input or calibrated scene geometry and lighting. In this work, we leverage a flash/no-flash image pair to analyze and edit scene illuminants based on their spectral differences. We derive a novel physics-based relationship between color variations in the observed flash/

no-flash intensities and the spectra and surface shading corresponding to individual scene illuminants. Our technique uses this constraint to automatically separate an image into constituent images lit by each illuminant. This separation can be used to support applications like white balancing, lighting editing, and RGB photometric stereo, where we demonstrate results that outperform state-of-the-art techniques on a wide range of images.

****

Trapping Light for Time of Flight

Ruilin Xu, Mohit Gupta, Shree K. Nayar; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6219-6227

We propose a novel imaging method for near-complete, surround, 3D reconstruction of geometrically complex objects, in a single shot. The key idea is to augment a time-of-flight (ToF) based 3D sensor with a multi-mirror system, called a light-trap. The shape of the trap is chosen so that light rays entering it bounce multiple times inside the trap, thereby visiting every position inside the trap multiple times from various directions. We show via simulations that this enables light rays to reach more than 99.9% of the surface of objects placed inside the trap, even those with strong occlusions, for example, lattice-shaped objects. The ToF sensor provides the path length for each light ray, which, along with the known shape of the trap, is used to reconstruct the complete paths of all the rays. This enables performing dense, surround 3D reconstructions of objects with highly complex 3D shapes, in a single shot. We have developed a proof-of-concept hardware prototype consisting of a pulsed ToF sensor, and a light trap built with planar mirrors. We demonstrate the effectiveness of the light trap based 3D reconstruction method on a variety of objects with a broad range of geometry and reflectance properties.

****

The Perception-Distortion Tradeoff

Yochai Blau, Tomer Michaeli; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6228-6237

Image restoration algorithms are typically evaluated by some distortion measure (e.g. PSNR, SSIM, IFC, VIF) or by human opinion scores that quantify perceived perceptual quality. In this paper, we prove mathematically that distortion and perceptual quality are at odds with each other. Specifically, we study the optimal probability for correctly discriminating the outputs of an image restoration algorithm from real images. We show that as the mean distortion decreases, this probability must increase (indicating worse perceptual quality). As opposed to the common belief, this result holds true for any distortion measure, and is not only a problem of the PSNR or SSIM criteria. However, as we show experimentally, for some measures it is less severe (e.g. distance between VGG features). We also show that generative-adversarial-nets (GANs) provide a principled way to approach the perception-distortion bound. This constitutes theoretical support to their observed success in low-level vision tasks. Based on our analysis, we propose a new methodology for evaluating image restoration methods, and use it to perform an extensive comparison between recent super-resolution algorithms.

****

Label Denoising Adversarial Network (LDAN) for Inverse Lighting of Faces

Hao Zhou, Jin Sun, Yaser Yacoob, David W. Jacobs; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6238-6247

Lighting estimation from faces is an important task and has applications in many areas such as image editing, intrinsic image decomposition, and image forgery detection. We propose to train a deep Convolutional Neural Network (CNN) to regress lighting parameters from a single face image. Lacking massive ground truth lighting labels for face images in the wild, we use an existing method to estimate lighting parameters, which are treated as ground truth with noise. To alleviate the effect of such noise, we utilize the idea of Generative Adversarial Networks (GAN) and propose a Label Denoising Adversarial Network (LDAN). LDAN makes use of synthetic data with accurate ground truth to help train a deep CNN for lighting regression on real face images. Experiments show that our network outperforms existing methods in producing consistent lighting parameters of different face

s under similar lighting conditions. To further evaluate the proposed method, we also apply it to regress object 2D key points where ground truth labels are available. Our experiments demonstrate its effectiveness on this application.
****

Optimal Structured Light à La Carte
Parsa Mirdehghan, Wenzheng Chen, Kiriakos N. Kutulakos; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6248-6257
We consider the problem of automatically generating sequences of structured-light patterns for active stereo triangulation of a static scene. Unlike existing approaches that use predetermined patterns and reconstruction algorithms tied to them, we generate patterns on the fly in response to generic specifications: number of patterns, projector-camera arrangement, workspace constraints, spatial frequency content, etc. Our pattern sequences are specifically optimized to minimize the expected rate of correspondence errors under those specifications for an unknown scene, and are coupled to a sequence-independent algorithm for per-pixel disparity estimation. To achieve this, we derive an objective function that is easy to optimize and follows from first principles within a maximum-likelihood framework. By minimizing it, we demonstrate automatic discovery of pattern sequences, in under three minutes on a laptop, that can outperform state-of-the-art triangulation techniques.
****

Tracking Multiple Objects Outside the Line of Sight Using Speckle Imaging
Brandon M. Smith, Matthew O'Toole, Mohit Gupta; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6258-6266
This paper presents techniques for tracking non-line-of-sight (NLOS) objects using speckle imaging. We develop a novel speckle formation and motion model where both the sensor and the source view objects only indirectly via a diffuse wall. We show that this NLOS imaging scenario is analogous to direct LOS imaging with the wall acting as a virtual, bare (lens-less) sensor. This enables tracking of a single, rigidly moving NLOS object using existing speckle-based motion estimation techniques. However, when imaging multiple NLOS objects, the speckle components due to different objects are superimposed on the virtual bare sensor image, and cannot be analyzed separately for recovering the motion of individual objects. We develop a novel clustering algorithm based on the statistical and geometrical properties of speckle images, which enables identifying the motion trajectories of multiple, independently moving NLOS objects. We demonstrate, for the first time, tracking individual trajectories of multiple objects around a corner with extreme precision (< 10 microns) using only off-the-shelf imaging components.
****

Inferring Light Fields From Shadows
Manel Baradad, Vickie Ye, Adam B. Yedidia, Frédo Durand, William T. Freeman, Gregory W. Wornell, Antonio Torralba; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6267-6275
We present a method for inferring a 4D light field of a hidden scene from 2D shadows cast by a known occluder on a diffuse wall. We do this by determining how light naturally reflected off surfaces in the hidden scene interacts with the occluder. By modeling the light transport as a linear system, and incorporating prior knowledge about light field structures, we can invert the system to recover the hidden scene. We demonstrate results of our inference method across simulations and experiments with different types of occluders. For instance, using the shadow cast by a real house plant, we are able to recover low resolution light fields with different levels of texture and parallax complexity. We provide two experimental results: a human subject and two planar elements at different depths.
****

Modifying Non-Local Variations Across Multiple Views
Tal Tlusty, Tomer Michaeli, Tali Dekel, Lihi Zelnik-Manor; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6276-6285
We present an algorithm for modifying small non-local variations between repeati

ng structures and patterns in multiple images of the same scene. The modification is consistent across views, even-though the images could have been photographed from different view points and under different lighting conditions. We show that when modifying each image independently the correspondence between them breaks and the geometric structure of the scene gets distorted. Our approach modifies the views while maintaining correspondence, hence, we succeed in modifying appearance and structure variations consistently. We demonstrate our methods on a number of challenging examples, photographed in different lighting, scales and view points.

****

Robust Video Content Alignment and Compensation for Rain Removal in a CNN Framework

Jie Chen, Cheen-Hau Tan, Junhui Hou, Lap-Pui Chau, He Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6286-6295

Rain removal is important for improving the robustness of outdoor vision based systems. Current rain removal methods show limitations either for complex dynamic scenes shot from fast moving cameras, or under torrential rain fall with opaque occlusions. We propose a novel derain algorithm, which applies superpixel (SP) segmentation to decompose the scene into depth consistent units. Alignment of scene contents are done at the SP level, which proves to be robust towards rain occlusion and fast camera motion. Two alignment output tensors, i.e., optimal temporal match tensor and sorted spatial-temporal match tensor, provide informative clues for rain streak location and occluded background contents to generate an intermediate derain output. These tensors will be subsequently prepared as input features for a convolutional neural network to restore high frequency details to the intermediate output for compensation of misalignment blur. Extensive evaluations show that up to 5dB reconstruction PSNR advantage is achieved over state-of-the-art methods. Visual inspection shows that much cleaner rain removal is achieved especially for highly dynamic scenes with heavy and opaque rainfall from a fast moving camera.

****

SfSNet: Learning Shape, Reflectance and Illuminance of Faces `in the Wild'

Soumyadip Sengupta, Angjoo Kanazawa, Carlos D. Castillo, David W. Jacobs; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6296-6305

We present SfSNet, an end-to-end learning framework for producing an accurate decomposition of an unconstrained human face image into shape, reflectance and illuminance. SfSNet is designed to reflect a physical lambertian rendering model. SfSNet learns from a mixture of labeled synthetic and unlabeled real world images. This allows the network to capture low frequency variations from synthetic and high frequency details from real images through the photometric reconstruction loss. SfSNet consists of a new decomposition architecture with residual blocks that learns a complete separation of albedo and normal. This is used along with the original image to predict lighting. SfSNet produces significantly better quantitative and qualitative results than state-of-the-art methods for inverse rendering and independent normal and illumination estimation.

****

Deep Photo Enhancer: Unpaired Learning for Image Enhancement From Photographs With GANs

Yu-Sheng Chen, Yu-Ching Wang, Man-Hsin Kao, Yung-Yu Chuang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6306-6314

This paper proposes an unpaired learning method for image enhancement. Given a set of photographs with the desired characteristics, the proposed method learns a photo enhancer which transforms an input image into an enhanced image with those characteristics. The method is based on the framework of two-way generative adversarial networks (GANs) with several improvements. First, we augment the U-Net with global features and show that it is more effective. The global U-Net acts as the generator in our GAN model. Second, we improve Wasserstein GAN (WGAN) wi th an adaptive weighting scheme. With this scheme, training converges faster and better, and is less sensitive to parameters than WGAN-GP. Finally, we propose to use individual batch normalization layers for generators in two-way GANs. It helps generators better adapt to their own input distributions. All together, they significantly improve the stability of GAN training for our application. Both quantitative and visual results show that the proposed method is effective for enhancing images.

****

LIME: Live Intrinsic Material Estimation

Abhimitra Meka, Maxim Maximov, Michael Zollhöfer, Avishek Chatterjee, Hans-Peter Seidel, Christian Richardt, Christian Theobalt; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6315-6324

We present the first end-to-end approach for real-time material estimation for general object shapes with uniform material that only requires a single color image as input. In addition to Lambertian surface properties, our approach fully automatically computes the specular albedo, material shininess, and a foreground segmentation. We tackle this challenging and ill-posed inverse rendering problem using recent advances in image-to-image translation techniques based on deep convolutional encoder–decoder architectures. The underlying core representations of our approach are specular shading, diffuse shading and mirror images, which allow to learn the effective and accurate separation of diffuse and specular albedo. In addition, we propose a novel highly efficient perceptual rendering loss that mimics real world image formation and obtains intermediate results even during run time. The estimation of material parameters at real-time frame rates enables exciting mixed reality applications, such as seamless illumination-consistent integration of virtual objects into realworld scenes, and virtual material cloning.We demonstrate our approach in a live setup, compare it to the state of the art, and demonstrate its effectiveness through quantitative and qualitative evaluation.

****

Learning to Detect Features in Texture Images

Linguang Zhang, Szymon Rusinkiewicz; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6325-6333

Local feature detection is a fundamental task in computer vision, and hand-crafted feature detectors such as SIFT have shown success in applications including image-based localization and registration. Recent work has used features detected in texture images for precise global localization, but is limited by the performance of existing feature detectors on textures, as opposed to natural images.

We propose an effective and scalable method for learning feature detectors for textures, which combines an existing "ranking" loss with an efficient fully-convolutional architecture as well as a new training-loss term that maximizes the "peakedness" of the response map. We demonstrate that our detector is more repeatable than existing methods, leading to improvements in a real-world texture-based localization application.

****

Learning to Extract a Video Sequence From a Single Motion-Blurred Image

Meiguang Jin, Givi Meishvili, Paolo Favaro; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6334-6342

We present a method to extract a video sequence from a single motion-blurred image. Motion-blurred images are the result of an averaging process, where instant frames are accumulated over time during the exposure of the sensor. Unfortunately, reversing this process is nontrivial. Firstly, averaging destroys the temporal ordering of the frames. Secondly, the recovery of a single frame is a blind deconvolution task, which is highly ill-posed. We present a deep learning scheme that gradually reconstructs a temporal ordering by sequentially extracting pairs of frames. Our main contribution is to introduce loss functions invariant to the temporal order. This lets a neural network choose during training what frame to output among the possible combinations. We also address the ill-posedness of deblurring by designing a network with a large receptive field and implemented via resampling to achieve a higher computational efficiency. Our proposed method can successfully retrieve sharp image sequences from a single motion blurred image and can generalize well on synthetic and real datasets captured with different cameras.

****

Lose the Views: Limited Angle CT Reconstruction via Implicit Sinogram Completion

Rushil Anirudh, Hyojin Kim, Jayaraman J. Thiagarajan, K. Aditya Mohan, Kyle Champley, Timo Bremer; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6343-6352

Computed Tomography (CT) reconstruction is a fundamental component to a wide variety of applications ranging from security, to healthcare. The classical techniques require measuring projections, called sinograms, from a full 180 degree view of the object. However, obtaining a full-view is not always feasible, such as when scanning irregular objects that limit flexibility of scanner rotation. The resulting limited angle sinograms are known to produce highly artifact-laden reconstructions with existing techniques. In this paper, we propose to address this problem using CTNet -- a system of 1D and 2D convolutional neural networks, that operates directly on a limited angle sinogram to predict the reconstruction. We use the x-ray transform on this prediction to obtain a ``completed'' sinogram, as if it came from a full 180 degree view. We feed this to standard analytical and iterative reconstruction techniques to obtain the final reconstruction. We show with extensive experimentation on a challenging real world dataset that this combined strategy outperforms many competitive baselines. We also propose a measure of confidence for the reconstruction that enables a practitioner to gauge the reliability of a prediction made by CTNet. We show that this measure is a strong indicator of quality as measured by the PSNR, while not requiring ground truth at test time. Finally, using a segmentation experiment, we show that our reconstruction also preserves the 3D structure of objects better than existing solutions.

****

A Common Framework for Interactive Texture Transfer

Yifang Men, Zhouhui Lian, Yingmin Tang, Jianguo Xiao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6353-6362

In this paper, we present a general-purpose solution to interactive texture transfer problems that better preserves both local structure and visual richness. It is challenging due to the diversity of tasks and the simplicity of required user guidance. The core idea of our common framework is to use multiple custom channels to dynamically guide the synthesis process. For interactivity, users can control the spatial distribution of stylized textures via semantic channels. The structure guidance, acquired by two stages of automatic extraction and propagation of structure information, provides a prior for initialization and preserves the salient structure by searching the nearest neighbor fields (NNF) with structure coherence. Meanwhile, texture coherence is also exploited to maintain similar style with the source image. In addition, we leverage an improved PatchMatch with extended NNF and matrix operations to obtain transformable source patches with richer geometric information at high speed. We demonstrate the effectiveness and superiority of our method on a variety of scenes through extensive comparisons with state-of-the-art algorithms.

****

AMNet: Memorability Estimation With Attention

Jiri Fajtl, Vasileios Argyriou, Dorothy Monekosso, Paolo Remagnino; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6363-6372

In this paper we present the design and evaluation of an end to end trainable, deep neural network with a visual attention mechanism for memorability estimation in still images. We analyze the suitability of transfer learning of deep models from image classification to the memorability task. Further on we study the impact of the attention mechanism on the memorability estimation and evaluate our network on the SUN Memorability and the LaMem dataset, the only large dataset with memorability labels to this date. Our network outperforms the existing state of the art models on both, the LaMem and SUN datasets in the term of the Spearman

’s rank correlation as well as mean squared error, approaching human consistency.

****

Blind Predicting Similar Quality Map for Image Quality Assessment

Da Pan, Ping Shi, Ming Hou, Zefeng Ying, Sizhe Fu, Yuan Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6373-6382

A key problem in blind image quality assessment (BIQA) is how to effectively model the properties of human visual system in a data-driven manner. In this paper, we propose a simple and efficient BIQA model based on a novel framework which consists of a fully convolutional neural network (FCNN) and a pooling network to solve this problem. In principle, FCNN is capable of predicting a pixel-by-pixel similar quality map only from a distorted image by using the intermediate similarity maps derived from conventional full-reference image quality assessment methods. The predicted pixel-by-pixel quality maps have good consistency with the distortion correlations between the reference and distorted images. Finally, a deep pooling network regresses the quality map into a score. Experiments have demonstrated that our predictions outperform many state-of-the-art BIQA methods.

****

Deep End-to-End Time-of-Flight Imaging

Shuochen Su, Felix Heide, Gordon Wetzstein, Wolfgang Heidrich; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6383-6392

We present an end-to-end image processing framework for time-of-flight (ToF) cameras. Existing ToF image processing pipelines consist of a sequence of operations including modulated exposures, denoising, phase unwrapping and multipath interference correction. While this cascaded modular design offers several benefits, such as closed-form solutions and power-efficient processing, it also suffers from error accumulation and information loss as each module can only observe the output from its direct predecessor, resulting in erroneous depth estimates. We depart from a conventional pipeline model and propose a deep convolutional neural network architecture that recovers scene depth directly from dual-frequency, raw ToF correlation measurements. To train this network, we simulate ToF images for a variety of scenes using a time-resolved renderer, devise depth-specific losses, and apply normalization and augmentation strategies to generalize this model to real captures. We demonstrate that the proposed network can efficiently exploit the spatio-temporal structures of ToF frequency measurements, and validate the performance of the joint multipath removal, denoising and phase unwrapping method on a wide range of challenging scenes.

****

Aperture Supervision for Monocular Depth Estimation

Pratul P. Srinivasan, Rahul Garg, Neal Wadhwa, Ren Ng, Jonathan T. Barron; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6393-6401

We present a novel method to train machine learning algorithms to estimate scene depths from a single image, by using the information provided by a camera's aperture as supervision. Prior works use a depth sensor's outputs or images of the same scene from alternate viewpoints as supervision, while our method instead uses images from the same viewpoint taken with a varying camera aperture. To enable learning algorithms to use aperture effects as supervision, we introduce two differentiable aperture rendering functions that use the input image and predicted depths to simulate the depth-of-field effects caused by real camera apertures. We train a monocular depth estimation network end-to-end to predict the scene depths that best explain these finite aperture images as defocus-blurred renderings of the input all-in-focus image.

****

Seeing Temporal Modulation of Lights From Standard Cameras

Naoki Sakakibara, Fumihiko Sakaue, Jun Sato; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6402-6410

In this paper, we propose a novel method for measuring the temporal modulation o f lights by using off-the-shelf cameras. In particular, we show that the invisible flicker patterns of various lights such as fluorescent lights can be measured by a simple combination of an off-the-shelf camera and any moving object with specular reflection. Unlike the existing methods, we do not need high speed cameras nor specially designed coded exposure cameras. Based on the extracted flicker patterns of environment lights, we also propose an efficient method for deblurring motion blurs in images. The proposed method enables us to deblur images with better frequency characteristics, which are induced by the flicker patterns of environment lights. The real image experiments show the efficiency of the proposed method.

****

Statistical Tomography of Microscopic Life

Aviad Levis, Yoav Y. Schechner, Ronen Talmon; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6411-6420

We achieve tomography of 3D volumetric natural objects, where each projected 2D image corresponds to a different specimen. Each specimen has unknown random 3D orientation, location, and scale. This imaging scenario is relevant to microscopic and mesoscopic organisms, aerosols and hydrosols viewed naturally by a microscope. In-class scale variation inhibits prior single-particle reconstruction methods. We thus generalize tomographic recovery to account for all degrees of freedom of a similarity transformation. This enables geometric self-calibration in imaging of transparent objects. We make the computational load manageable and reach good quality reconstruction in a short time. This enables extraction of statistics that are important for a scientific study of specimen populations, specifically size distribution parameters. We apply the method to study of plankton.

****

Divide and Conquer for Full-Resolution Light Field Deblurring

M. R. Mahesh Mohan, A. N. Rajagopalan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6421-6429

The increasing popularity of computational light field (LF) cameras has necessitated the need for tackling motion blur which is a ubiquitous phenomenon in hand-held photography. The state-of-the-art method for blind deblurring of LFs of general 3D scenes is limited to handling only downsampled LF, both in spatial and angular resolution. This is due to the computational overhead involved in processing data-hungry full-resolution 4D LF altogether. Moreover, the method warrants high-end GPUs for optimization and is ineffective for wide-angle settings and irregular camera motion. In this paper, we introduce a new blind motion deblurring strategy for LFs which alleviates these limitations significantly. Our model achieves this by isolating 4D LF motion blur across the 2D subaperture images, thus paving the way for independent deblurring of these subaperture images. Furthermore, our model accommodates common camera motion parameterization across the subaperture images. Consequently, blind deblurring of any single subaperture image elegantly paves the way for cost-effective non-blind deblurring of the other subaperture images. Our approach is CPU-efficient computationally and can effectively deblur full-resolution LFs.

****

Multispectral Image Intrinsic Decomposition via Subspace Constraint

Qian Huang, Weixin Zhu, Yang Zhao, Linsen Chen, Yao Wang, Tao Yue, Xun Cao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6430-6439

Multispectral images contain many clues of surface characteristics of the objects, thus can be used in many computer vision tasks, e.g., recolorization and segmentation. However, due to the complex geometry structure of natural scenes, the spectra curves of the same surface can look very different under different illuminations and from different angles. In this paper, a new Multispectral Image Intrinsic Decomposition model (MIID) is presented to decompose the shading and reflectance from a single multispectral image. We extend the Retinex model, which is proposed for RGB image intrinsic decomposition, for multispectral domain. Based on this, a subspace constraint is introduced to both the shading and reflectance spectral space to reduce the ill-posedness of the problem and make the problem solvable. A dataset of 22 scenes is given with the ground truth of shadings and reflectance to facilitate objective evaluations. The experiments demonstrate the effectiveness of the proposed method.

****

Improving Color Reproduction Accuracy on Cameras

Hakki Can Karaimer, Michael S. Brown; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6440-6449

One of the key operations performed on a digital camera is to map the sensor-specific color space to a standard perceptual color space.   This procedure involves the application of a white-balance correction followed by a color space transform. The current approach for this colorimetric mapping is based on an interpolation of pre-calibrated color space transforms computed for two fixed illuminations (i.e., two white-balance settings). Images captured under different illuminations are subject to less color accuracy due to the use of this interpolation process.   In this paper, we discuss the limitations of the current colorimetric mapping approach and propose two methods that are able to improve color accuracy. We evaluate our approach on seven different cameras and show improvements of up to 30% (DSLR cameras) and 59% (mobile phone cameras) in terms of color reproduction error.

****

A Closer Look at Spatiotemporal Convolutions for Action Recognition

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, Manohar Paluri; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6450-6459

In this paper we discuss several forms of spatiotemporal convolutions for video analysis and study their effects on action recognition. Our motivation stems from the observation that 2D CNNs applied to individual frames of the video have remained solid performers in action recognition. In this work we empirically demonstrate the accuracy advantages of 3D CNNs over 2D CNNs within the framework of residual learning. Furthermore, we show that factorizing the 3D convolutional filters into separate spatial and temporal components yields significantly gains in accuracy. Our empirical study leads to the design of a new spatiotemporal convolutional block ``R(2+1)D'' which produces CNNs that achieve results comparable or superior to the state-of-the-art on Sports-1M, Kinetics, UCF101, and HMDB51.

****

Inferring Shared Attention in Social Scene Videos

Lifeng Fan, Yixin Chen, Ping Wei, Wenguan Wang, Song-Chun Zhu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6460-6468

This paper addresses a new problem of inferring shared attention in third-person social scene videos. Shared attention is a phenomenon that two or more individuals simultaneously look at a common target in social scenes. Perceiving and identifying shared attention in videos plays crucial roles in social activities and social scene understanding. We propose a spatial-temporal neural network to detect shared attention intervals in videos and predict shared attention locations in frames. In each video frame, human gaze directions and potential target boxes are two key features for spatially detecting shared attention in the social scene. In temporal domain, a convolutional Long Short- Term Memory network utilizes the temporal continuity and transition constraints to optimize the predicted shared attention heatmap. We collect a new dataset VideoCoAtt from public TV show videos, containing 380 complex video sequences with more than 492,000 frames that include diverse social scenes for shared attention study. Experiments on this dataset show that our model can effectively infer shared attention in videos. We also empirically verify the effectiveness of different components in our model.

****

Making Convolutional Networks Recurrent for Visual Sequence Learning

Xiaodong Yang, Pavlo Molchanov, Jan Kautz; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6469-6478

Recurrent neural networks (RNNs) have emerged as a powerful model for a broad range of machine learning problems that involve sequential data. While an abundanc

e of work exists to understand and improve RNNs in the context of language and audio signals such as language modeling and speech recognition, relatively little attention has been paid to analyze or modify RNNs for visual sequences, which by nature have distinct properties. In this paper, we aim to bridge this gap and present the first large-scale exploration of RNNs for visual sequence learning. In particular, with the intention of leveraging the strong generalization capacity of pre-trained convolutional neural networks (CNNs), we propose a novel and effective approach, PreRNN, to make pre-trained CNNs recurrent by transforming convolutional layers or fully connected layers into recurrent layers. We conduct extensive evaluations on three representative visual sequence learning tasks: sequential face alignment, dynamic hand gesture recognition, and action recognition. Our experiments reveal that PreRNN consistently outperforms the traditional RNNs and achieves state-of-the-art results on the three applications, suggesting that PreRNN is more suitable for visual sequence learning.
****

Real-World Anomaly Detection in Surveillance Videos
Waqas Sultani, Chen Chen, Mubarak Shah; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6479-6488

Surveillance videos are able to capture a variety of realistic anomalies. In this paper, we propose to learn anomalies by exploiting both normal and anomalous videos. To avoid annotating the anomalous segments or clips in training videos, which is very time consuming, we propose to learn anomaly through the deep multiple instance ranking framework by leveraging weakly labeled training videos, ie the training labels (anomalous or normal) are at video-level instead of clip-level. In our approach, we consider normal and anomalous videos as bags and video segments as instances in multiple instance learning (MIL), and automatically learn a deep anomaly ranking model that predicts high anomaly scores for anomalous video segments. Furthermore, we introduce sparsity and temporal smoothness constraints in the ranking loss function to better localize anomaly during training. We also introduce a new large-scale first of its kind dataset of 128 hours of videos. It consists of 1900 long and untrimmed real-world surveillance videos, with 13 realistic anomalies such as fighting, road accident, burglary, robbery, etc. as well as normal activities. This dataset can be used for two tasks. First, general anomaly detection considering all anomalies in one group and all normal activities in another group. Second, for recognizing each of 13 anomalous activities. Our experimental results show that our MIL method for anomaly detection achieves significant improvement on anomaly detection performance as compared to the state-of-the-art approaches. We provide the results of several recent deep learning baselines on anomalous activity recognition. The low recognition performance of these baselines reveals that our dataset is very challenging and opens more opportunities for future work.
****

Viewpoint-Aware Attentive Multi-View Inference for Vehicle Re-Identification
Yi Zhou, Ling Shao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6489-6498

Vehicle re-identification (re-ID) has the huge potential to contribute to the intelligent video surveillance. However, it suffers from challenges that different vehicle identities with a similar appearance have little inter-instance discrepancy while one vehicle usually has large intra-instance differences under viewpoint and illumination variations. Previous methods address vehicle re-ID by simply using visual features from originally captured views and usually exploit the spatial-temporal information of the vehicles to refine the results. In this paper, we propose a Viewpoint-aware Attentive Multi-view Inference (VAMI) model that only requires visual information to solve the multi-view vehicle re-ID problem. Given vehicle images of arbitrary viewpoints, the VAMI extracts the single-view feature for each input image and aims to transform the features into a global multi-view feature representation so that pairwise distance metric learning can be better optimized in such a viewpoint-invariant feature space. The VAMI adopts a viewpoint-aware attention model to select core regions at different viewpoints and implement effective multi-view feature inference by an adversarial training architecture. Extensive experiments validate the effectiveness of each proposed component and illustrate that our approach achieves consistent improvements over state-of-the-art vehicle re-ID methods on two public datasets: VeRi and Vehicle ID.

****

Efficient Video Object Segmentation via Network Modulation

Linjie Yang, Yanran Wang, Xuehan Xiong, Jianchao Yang, Aggelos K. Katsaggelos; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6499-6507

Video object segmentation targets segmenting a specific object throughout a video sequence when given only an annotated first frame. Recent deep learning based approaches find it effective to fine-tune a general-purpose segmentation model on the annotated frame using hundreds of iterations of gradient descent. Despite the high accuracy that these methods achieve, the fine-tuning process is inefficient and fails to meet the requirements of real world applications. We propose a novel approach that uses a single forward pass to adapt the segmentation model to the appearance of a specific object. Specifically, a second meta neural network named modulator is trained to manipulate the intermediate layers of the segmentation network given limited visual and spatial information of the target object. The experiments show that our approach is 70 times faster than fine-tuning approaches and achieves similar accuracy.

****

Weakly-Supervised Action Segmentation With Iterative Soft Boundary Assignment

Li Ding, Chenliang Xu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6508-6516

In this work, we address the task of weakly-supervised human action segmentation in long, untrimmed videos. Recent methods have relied on expensive learning models, such as Recurrent Neural Networks (RNN) and Hidden Markov Models (HMM). However, these methods suffer from expensive computational cost, thus are unable to be deployed in large scale. To overcome the limitations, the keys to our design are efficiency and scalability. We propose a novel action modeling framework, which consists of a new temporal convolutional network, named Temporal Convolutional Feature Pyramid Network (TCFPN), for predicting frame-wise action labels, and a novel training strategy for weakly-supervised sequence modeling, named Iterative Soft Boundary Assignment (ISBA), to align action sequences and update the network in an iterative fashion. The proposed framework is evaluated on two benchmark datasets, Breakfast and Hollywood Extended, with four different evaluation metrics. Extensive experimental results show that our methods achieve competitive or superior performance to state-of-the-art methods.

****

Depth-Aware Stereo Video Retargeting

Bing Li, Chia-Wen Lin, Boxin Shi, Tiejun Huang, Wen Gao, C.-C. Jay Kuo; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6517-6525

As compared with traditional video retargeting, stereo video retargeting poses new challenges because stereo video contains the depth information of salient objects and its time dynamics. In this work, we propose a depth-aware stereo video retargeting method by imposing the depth fidelity constraint. The proposed depth-aware retargeting method reconstructs the 3D scene to obtain the depth information of salient objects. We cast it as a constrained optimization problem, where the total cost function includes the shape, temporal and depth distortions of salient objects. As a result, the solution can preserve the shape, temporal and depth fidelity of salient objects simultaneously. It is demonstrated by experimental results that the depth-aware retargeting method achieves higher retargeting quality and provides better user experience.

****

Instance Embedding Transfer to Unsupervised Video Object Segmentation

Siyang Li, Bryan Seybold, Alexey Vorobyov, Alireza Fathi, Qin Huang, C.-C. Jay Kuo; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6526-6535

We propose a method for unsupervised video object segmentation by transferring the knowledge encapsulated in image-based instance embedding networks. The instance embedding network produces an embedding vector for each pixel that enables identifying all pixels belonging to the same object. Though trained on static images, the instance embeddings are stable over consecutive video frames, which allows us to link objects together over time. Thus, we adapt the instance networks trained on static images to video object segmentation and incorporate the embeddings with objectness and optical flow features, without model retraining or online fine-tuning. The proposed method outperforms state-of-the-art unsupervised segmentation methods in the DAVIS dataset and the FBMS dataset.
****

Future Frame Prediction for Anomaly Detection – A New Baseline
Wen Liu, Weixin Luo, Dongze Lian, Shenghua Gao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6536-6545

Anomaly detection in videos refers to the identification of events that do not conform to expected behavior. However, almost all existing methods tackle the problem by minimizing the reconstruction errors of training data, which cannot guarantee a larger reconstruction error for an abnormal event. In this paper, we propose to tackle the anomaly detection problem within a video prediction framework. To the best of our knowledge, this is the first work that leverages the difference between a predicted future frame and its ground truth to detect an abnormal event. To predict a future frame with higher quality for normal events, other than the commonly used appearance (spatial) constraints on intensity and gradient, we also introduce a motion (temporal) constraint in video prediction by enforcing the optical flow between predicted frames and ground truth frames to be consistent, and this is the first work that introduces a temporal constraint into the video prediction task. Such spatial and motion constraints facilitate the future frame prediction for normal events, and consequently facilitate to identify those abnormal events that do not conform the expectation. Extensive experiments on both a toy dataset and some publicly available datasets validate the effectiveness of our method in terms of robustness to the uncertainty in normal events and the sensitivity to abnormal events.
****

Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?
Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6546-6555

The purpose of this study is to determine whether current video datasets have sufficient data for training very deep convolutional neural networks (CNNs) with spatio-temporal three-dimensional (3D) kernels. Recently, the performance levels of 3D CNNs in the field of action recognition have improved significantly. However, to date, conventional research has only explored relatively shallow 3D architectures. We examine the architectures of various 3D CNNs from relatively shallow to very deep ones on current video datasets. Based on the results of those experiments, the following conclusions could be obtained: (i) ResNet-18 training resulted in significant overfitting for UCF-101, HMDB-51, and ActivityNet but not for Kinetics. (ii) The Kinetics dataset has sufficient data for training of deep 3D CNNs, and enables training of up to 152 ResNets layers, interestingly similar to 2D ResNets on ImageNet. ResNeXt-101 achieved 78.4% average accuracy on the Kinetics test set. (iii) Kinetics pretrained simple 3D architectures outperforms complex 2D architectures, and the pretrained ResNeXt-101 achieved 94.5% and 70.2% on UCF-101 and HMDB-51, respectively. The use of 2D CNNs trained on ImageNet has produced significant progress in various tasks in image. We believe that using deep 3D CNNs together with Kinetics will retrace the successful history of 2D CNNs and ImageNet, and stimulate advances in computer vision for videos. The codes and pretrained models used in this study are publicly available. https://github.com/kenshohara/3D-ResNets-PyTorch
****

Dynamic Video Segmentation Network
Yu-Syuan Xu, Tsu-Jui Fu, Hsuan-Kung Yang, Chun-Yi Lee; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6556-6565

In this paper, we present a detailed design of dynamic video segmentation network (DVSNet) for fast and efficient semantic video segmentation. DVSNet consists of two convolutional neural networks: a segmentation network and a flow network. The former generates highly accurate semantic segmentations, but is deeper and slower. The latter is much faster than the former, but its output requires further processing to generate less accurate semantic segmentations. We explore the use of a decision network to adaptively assign different frame regions to different networks based on a metric called expected confidence score. Frame regions with a higher expected confidence score traverse the flow network. Frame regions with a lower expected confidence score have to pass through the segmentation network. We have extensively performed experiments on various configurations of DVSNet, and investigated a number of variants for the proposed decision network. The experimental results show that our DVSNet is able to achieve up to 70.4% mIoU at 19.8 fps on the Cityscape dataset. A high speed version of DVSNet is able to deliver an fps of 30.4 with 63.2% mIoU on the same dataset. DVSNet is also able to reduce up to 95% of the computational workloads.

****

Recognize Actions by Disentangling Components of Dynamics

Yue Zhao, Yuanjun Xiong, Dahua Lin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6566-6575

Despite the remarkable progress in action recognition over the past several years, existing methods remain limited in efficiency and effectiveness. The methods treating appearance and motion as separate streams are usually subject to the cost of optical flow computation, while those relying on 3D convolution on the original video frames often yield inferior performance in practice. In this paper, we propose a new ConvNet architecture for video representation learning, which can derive disentangled components of dynamics purely from raw video frames, without the need of optical flow estimation. Particularly, the learned representation comprises three components for representing static appearance, apparent motion, and appearance changes. We introduce 3D pooling, cost volume processing, and warped feature differences, respectively for extracting the three components above. These modules are incorporated as three branches in our unified network, which share the underlying features and are learned jointly in an end-to-end manner. On two large datasets UCF101 and Kinetics our method obtained competitive performances with high efficiency, using only the RGB frame sequence as input.

****

Motion-Appearance Co-Memory Networks for Video Question Answering

Jiyang Gao, Runzhou Ge, Kan Chen, Ram Nevatia; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6576-6585

Video Question Answering (QA) is an important task in understanding video temporal structure. We observe that there are three unique attributes of video QA compared with image QA: (1) it deals with long sequences of images containing richer information not only in quantity but also in variety; (2) motion and appearance information are usually correlated with each other and able to provide useful attention cues to the other; (3) different questions require different number of frames to infer the answer. Based these observations, we propose a motion-appearance co-memory network for video QA. Our networks are built on concepts from Dynamic Memory Network (DMN) and introduces new mechanisms for video QA. Specifically, there are three salient aspects: (1) a co-memory attention mechanism that utilizes cues from both motion and appearance to generate attention; (2) a temporal conv-deconv network to generate multi-level contextual facts; (3) a dynamic fact ensemble method to construct temporal representation dynamically for different questions. We evaluate our method on TGIF-QA dataset, and the results outperform state-of-the-art significantly on all four tasks of TGIF-QA.

****

Learning to Understand Image Blur

Shanghang Zhang, Xiaohui Shen, Zhe Lin, Radomír M■ch, João P. Costeira, José M. F. Moura; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6586-6595

While many approaches have been proposed to estimate and remove blur in a photo, few efforts were made to have an algorithm automatically understand the blur desirability: whether the blur is desired or not, and how it affects the quality of the photo. Such a task not only relies on low-level visual features to identify blurry regions, but also requires high-level understanding of the image content as well as user intent during photo capture. In this paper, we propose a unified framework to estimate a spatially-varying blur map and understand its desirability in terms of image quality at the same time. In particular, we use a dilated fully convolutional neural network with pyramid pooling and boundary refinement layers to generate high-quality blur response maps. If blur exists, we classify its desirability to three levels ranging from good to bad, by distilling high-level semantics and learning an attention map to adaptively localize the important content in the image. The whole framework is end-to-end jointly trained with both supervisions of pixel-wise blur responses and image-wise blur desirability levels. Considering the limitations of existing image blur datasets, we collected a new large-scale dataset with both annotations to facilitate training. The proposed methods are extensively evaluated on two datasets and demonstrate state-of-the-art performance on both tasks.

****

Dense Decoder Shortcut Connections for Single-Pass Semantic Segmentation
Piotr Bilinski, Victor Prisacariu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6596-6605

We propose a novel end-to-end trainable, deep, encoder-decoder architecture for single-pass semantic segmentation. Our approach is based on a cascaded architecture with feature-level long-range skip connections. The encoder incorporates the structure of ResNeXt's residual building blocks and adopts the strategy of repeating a building block that aggregates a set of transformations with the same topology. The decoder features a novel architecture, consisting of blocks, that (i) capture context information, (ii) generate semantic features, and (iii) enable fusion between different output resolutions. Crucially, we introduce dense decoder shortcut connections to allow decoder blocks to use semantic feature maps from all previous decoder levels, i.e. from all higher-level feature maps. The dense decoder connections allow for effective information propagation from one decoder block to another, as well as for multi-level feature fusion that significantly improves the accuracy. Importantly, these connections allow our method to obtain state-of-the-art performance on several challenging datasets, without the need of time-consuming multi-scale averaging of previous works.

****

Generative Adversarial Image Synthesis With Decision Tree Latent Controller
Takuhiro Kaneko, Kaoru Hiramatsu, Kunio Kashino; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6606-6615

This paper proposes the decision tree latent controller generative adversarial network (DTLC-GAN), an extension of a GAN that can learn hierarchically interpretable representations without relying on detailed supervision. To impose a hierarchical inclusion structure on latent variables, we incorporate a new architecture called the DTLC into the generator input. The DTLC has a multiple-layer tree structure in which the ON or OFF of the child node codes is controlled by the parent node codes. By using this architecture hierarchically, we can obtain the latent space in which the lower layer codes are selectively used depending on the higher layer ones. To make the latent codes capture salient semantic features of images in a hierarchically disentangled manner in the DTLC, we also propose a hierarchical conditional mutual information regularization and optimize it with a newly defined curriculum learning method that we propose as well. This makes it possible to discover hierarchically interpretable representations in a layer-by-layer manner on the basis of information gain by only using a single DTLC-GAN model. We evaluated the DTLC-GAN on various datasets, i.e., MNIST, CIFAR-10, Tiny ImageNet, 3D Faces, and CelebA, and confirmed that the DTLC-GAN can learn hierarchically interpretable representations with either unsupervised or weakly supervised settings. Furthermore, we applied the DTLC-GAN to image-retrieval tasks and showed its effectiveness in representation learning.

****

Learning a Discriminative Prior for Blind Image Deblurring

Lerenhan Li, Jinshan Pan, Wei-Sheng Lai, Changxin Gao, Nong Sang, Ming-Hsuan Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6616-6625

We present an effective blind image deblurring method based on a data-driven discriminative prior. Our work is motivated by the fact that a good image prior should favor clear images over blurred images. To obtain such an image prior for deblurring, we formulate the image prior as a binary classifier which can be achieved by a deep convolutional neural network (CNN). The learned image prior has a significant discriminative property and is able to distinguish whether the image is clear or not. Embedded into the maximum a posterior (MAP) framework, it helps blind deblurring on various scenarios, including natural, face, text, and low-illumination images. However, it is difficult to optimize the deblurring method with the learned image prior as it involves a non-linear CNN. Therefore, we develop an efficient numerical approach based on the half-quadratic splitting method and gradient decent algorithm to solve the proposed model. Furthermore, the proposed model can be easily extended to non-uniform deblurring. Both qualitative and quantitative experimental results show that our method performs favorably against state-of-the-art algorithms as well as domain-specific image deblurring approaches.

****

Frame-Recurrent Video Super-Resolution

Mehdi S. M. Sajjadi, Raviteja Vemulapalli, Matthew Brown; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6626-6634

Recent advances in video super-resolution have shown that convolutional neural networks combined with motion compensation are able to merge information from multiple low-resolution (LR) frames to generate high-quality images. Current state-of-the-art methods process a batch of LR frames to generate a single high-resolution (HR) frame and run this scheme in a sliding window fashion over the entire video, effectively treating the problem as a large number of separate multi-frame super-resolution tasks. This approach has two main weaknesses: 1) Each input frame is processed and warped multiple times, increasing the computational cost, and 2) each output frame is estimated independently conditioned on the input frames, limiting the system's ability to produce temporally consistent results. In this work, we propose an end-to-end trainable frame-recurrent video super-resolution framework that uses the previously inferred HR estimate to super-resolve the subsequent frame. This naturally encourages temporally consistent results and reduces the computational cost by warping only one image in each step. Furthermore, due to its recurrent nature, the proposed method has the ability to assimilate a large number of previous frames without increased computational demands. Extensive evaluations and comparisons with previous methods validate the strengths of our approach and demonstrate that the proposed framework is able to significantly outperform the current state of the art.

****

Discovering Point Lights With Intensity Distance Fields

Edward Zhang, Michael F. Cohen, Brian Curless; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6635-6643

We introduce the light localization problem. A scene is illuminated by a set of unobserved isotropic point lights. Given the geometry, materials, and illuminated appearance of the scene, the light localization problem is to completely recover the number, positions, and intensities of the lights. We first present a scene transform that identifies likely light positions. Based on this transform, we develop an iterative algorithm to locate remaining lights and determine all light intensities. We demonstrate the success of this method in a large set of 2D synthetic scenes, and show that it extends to 3D, in both synthetic scenes and real-world scenes.

****

Video Rain Streak Removal by Multiscale Convolutional Sparse Coding

Minghan Li, Qi Xie, Qian Zhao, Wei Wei, Shuhang Gu, Jing Tao, Deyu Meng; Proceed ings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6644-6653

Videos captured by outdoor surveillance equipments sometimes contain unexpected rain streaks, which brings difficulty in subsequent video processing tasks. Rain streak removal from a video is thus an important topic in recent computer vision research. In this paper, we raise two intrinsic characteristics specifically possessed by rain streaks. Firstly, the rain streaks in a video contain repetitive local patterns sparsely scattered over different positions of the video. Secondly, the rain streaks are with multiscale configurations due to their occurrence on positions with different distances to the cameras. Based on such understanding, we specifically formulate both characteristics into a multiscale convolutional sparse coding (MS-CSC) model for the video rain streak removal task. Specifically, we use multiple convolutional filters convolved on the sparse feature maps to deliver the former characteristic, and further use multiscale filters to represent different scales of rain streaks. Such a new encoding manner makes the proposed method capable of properly extracting rain streaks from videos, thus getting fine video deraining effects. Experiments implemented on synthetic and real videos verify the superiority of the proposed method, as compared with the state-of-the-art ones along this research line, both visually and quantitatively.

****

Stereoscopic Neural Style Transfer

Dongdong Chen, Lu Yuan, Jing Liao, Nenghai Yu, Gang Hua; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6654-6663

This paper presents the first attempt at stereoscopic neural style transfer, which responds to the emerging demand for 3D movies or AR/VR. We start with a careful examination of applying existing monocular style transfer methods to left and right views of stereoscopic images separately. This reveals that the original disparity consistency cannot be well preserved in the final stylization results, which causes 3D fatigue to the viewers. To address this issue, we incorporate a new disparity loss into the widely adopted style loss function by enforcing the bidirectional disparity constraint in non-occluded regions. For a practical real-time solution, we propose the first feed-forward network by jointly training a stylization sub-network and a disparity sub-network, and integrate them in a feature level middle domain. Our disparity sub-network is also the first end-to-end network for simultaneous bidirectional disparity and occlusion mask estimation. Finally, our network is effectively extended to stereoscopic videos, by considering both temporal coherence and disparity consistency. We will show that the proposed method clearly outperforms the baseline algorithms both quantitatively and qualitatively.

****

Multi-Frame Quality Enhancement for Compressed Video

Ren Yang, Mai Xu, Zulin Wang, Tianyi Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6664-6673

The past few years have witnessed great success in applying deep learning to enhance the quality of compressed image/video. The existing approaches mainly focus on enhancing the quality of a single frame, ignoring the similarity between consecutive frames. In this paper, we investigate that heavy quality fluctuation exists across compressed video frames, and thus low quality frames can be enhanced using the neighboring high quality frames, seen as Multi-Frame Quality Enhancement (MFQE). Accordingly, this paper proposes an MFQE approach for compressed video, as a first attempt in this direction. In our approach, we firstly develop a Support Vector Machine (SVM) based detector to locate Peak Quality Frames (PQFs) in compressed video. Then, a novel Multi-Frame Convolutional Neural Network (MF-CNN) is designed to enhance the quality of compressed video, in which the non-PQF and its nearest two PQFs are as the input. The MF-CNN compensates motion between the non-PQF and PQFs through the Motion Compensation subnet (MC-subnet). Subsequently, the Quality Enhancement subnet (QE-subnet) reduces compression artifacts of the non-PQF with the help of its nearest PQFs. Finally, the experiments v alidate the effectiveness and generality of our MFQE approach in advancing the state-of-the-art quality enhancement of compressed video. The code of our MFQE approach is available at https://github.com/ryangBUAA/MFQE.git.

****

CNN Based Learning Using Reflection and Retinex Models for Intrinsic Image Decomposition
Anil S. Baslamisli, Hoang-An Le, Theo Gevers; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6674-6683

Most of the traditional work on intrinsic image decomposition rely on deriving priors about scene characteristics. On the other hand, recent research use deep learning models as in-and-out black box and do not consider the well-established, traditional image formation process as the basis of their intrinsic learning process. As a consequence, although current deep learning approaches show superior performance when considering quantitative benchmark results, traditional approaches are still dominant in achieving high qualitative results. In this paper, the aim is to exploit the best of the two worlds. A method is proposed that (1) is empowered by deep learning capabilities, (2) considers a physics-based reflection model to steer the learning process, and (3) exploits the traditional approach to obtain intrinsic images by exploiting reflectance and shading gradient information. The proposed model is fast to compute and allows for the integration of all intrinsic components. To train the new model, an object centered large-scale datasets with intrinsic ground-truth images are created. The evaluation results demonstrate that the new model outperforms existing methods. Visual inspection shows that the image formation loss function augments color reproduction and the use of gradient information produces sharper edges. Datasets, models and higher resolution images are available at https://ivi.fnwi.uva.nl/cv/retinet.

****

Image Restoration by Estimating Frequency Distribution of Local Patches
Jaeyoung Yoo, Sang-ho Lee, Nojun Kwak; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6684-6692

In this paper, we propose a method to solve the image restoration problem, which tries to restore the details of a corrupted image, especially due to the loss caused by JPEG compression. We have treated an image in the frequency domain to explicitly restore the frequency components lost during image compression. In doing so, the distribution in the frequency domain is learned using the cross entropy loss. Unlike recent approaches, we have reconstructed the details of an image without using the scheme of adversarial training. Rather, the image restoration problem is treated as a classification problem to determine the frequency coefficient for each frequency band in an image patch. In this paper, we show that the proposed method effectively restores a JPEG-compressed image with more detailed high frequency components, making the restored image more vivid.

****

Latent RANSAC
Simon Korman, Roee Litman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6693-6702

We present a method that can evaluate a RANSAC hypothesis in constant time, i.e. independent of the size of the data. A key observation here is that correct hypotheses are tightly clustered together in the latent parameter domain. In a manner similar to the generalized Hough transform we seek to find this cluster, only that we need as few as two votes for a successful detection. Rapidly locating such pairs of similar hypotheses is made possible by adapting the recent "Random Grids" range-search technique. We only perform the usual (costly) hypothesis verification stage upon the discovery of a close pair of hypotheses. We show that this event rarely happens for incorrect hypotheses, enabling a significant speedup of the RANSAC pipeline. The suggested approach is applied and tested on three robust estimation problems: camera localization, 3D rigid alignment and 2D-homography estimation. We perform rigorous testing on both synthetic and real datasets, demonstrating an improvement in efficiency without a compromise in accuracy. Furthermore, we achieve state-of-the-art 3D alignment results on the challenging ``Redwood'' loop-closure challenge.

****

Two-Stream Convolutional Networks for Dynamic Texture Synthesis

Matthew Tesfaldet, Marcus A. Brubaker, Konstantinos G. Derpanis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6703-6712

We introduce a two-stream model for dynamic texture synthesis. Our model is based on pre-trained convolutional networks (ConvNets) that target two independent tasks: (i) object recognition, and (ii) optical flow prediction. Given an input dynamic texture, statistics of filter responses from the object recognition ConvNet encapsulate the per-frame appearance of the input texture, while statistics of filter responses from the optical flow ConvNet model its dynamics. To generate a novel texture, a randomly initialized input sequence is optimized to match the feature statistics from each stream of an example texture. Inspired by recent work on image style transfer and enabled by the two-stream model, we also apply the synthesis approach to combine the texture appearance from one texture with the dynamics of another to generate entirely novel dynamic textures. We show that our approach generates novel, high quality samples that match both the framewise appearance and temporal evolution of input texture. Finally, we quantitatively evaluate our texture synthesis approach with a thorough user study.

****

Towards Open-Set Identity Preserving Face Synthesis

Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, Gang Hua; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6713-6722

We propose a framework based on Generative Adversarial Networks to disentangle the identity and attributes of faces, such that we can conveniently recombine different identities and attributes for identity preserving face synthesis in open domains. Previous identity preserving face synthesis processes are largely confined to synthesizing faces with known identities that are already in the training dataset. To synthesize a face with identity outside the training dataset, our framework requires one input image of that subject to produce an identity vector, and any other input face image to extract an attribute vector capturing, e.g., pose, emotion, illumination, and even the background. We then recombine the identity vector and the attribute vector to synthesize a new face of the subject with the extracted attribute. Our proposed framework does not need to annotate the attributes of faces in any way. It is trained with an asymmetric loss function to better preserve the identity and stabilize the training process. It can also effectively leverage large amounts of unlabeled training face images to further improve the fidelity of the synthesized faces for subjects that are not presented in the labeled training face dataset. Our experiments demonstrate the efficacy of the proposed framework. We also present its usage in a much broader set of applications including face frontalization, face attribute morphing, and face adversarial example detection.

****

A Revised Underwater Image Formation Model

Derya Akkaynak, Tali Treibitz; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6723-6732

The current underwater image formation model descends from atmospheric dehazing equations where attenuation is a weak function of wavelength. We recently showed that this model introduces significant errors and dependencies in the estimation of the direct transmission signal because underwater, light attenuates in a wavelength-dependent manner. Here, we show that the backscattered signal derived from the current model also suffers from dependencies that were previously unaccounted for. In doing so, we use oceanographic measurements to derive the physically valid space of backscatter, and further show that the wideband coefficients that govern backscatter are different than those that govern direct transmission, even though the current model treats them to be the same. We propose a revised equation for underwater image formation that takes these differences into account, and validate it through in situ experiments underwater. This revised model might explain frequent instabilities of current underwater color reconstruction mo

dels, and calls for the development of new methods.
****

Graph-Cut RANSAC

Daniel Barath, Ji■í Matas; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6733-6741

A novel method for robust estimation, called Graph-Cut RANSAC, GC-RANSAC in short, is introduced. To separate inliers and outliers, it runs the graph-cut algorithm in the local optimization (LO) step which is applied when a so-far-the-best model is found. The proposed LO step is conceptually simple, easy to implement, globally optimal and efficient. GC-RANSAC is shown experimentally, both on synthesized tests and real image pairs, to be more geometrically accurate than state-of-the-art methods on a range of problems, e.g. line fitting, homography, affine transformation, fundamental and essential matrix estimation. It runs in real-time for many problems at a speed approximately equal to that of the less accurate alternatives (in milliseconds on standard CPU).
****

Temporal Deformable Residual Networks for Action Segmentation in Videos

Peng Lei, Sinisa Todorovic; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6742-6751

This paper is about temporal segmentation of human actions in videos. We introduce a new model -- temporal deformable residual network (TDRN) -- aimed at analyzing video intervals at multiple temporal scales for labeling video frames. Our TDRN computes two parallel temporal streams: i) Residual stream that analyzes video information at its full temporal resolution, and ii) Pooling/unpooling stream that captures long-range video information at different scales. The former facilitates local, fine-scale action segmentation, and the latter uses multiscale context for improving accuracy of frame classification. These two streams are computed by a set of temporal residual modules with deformable convolutions, and fused by temporal residuals at the full video resolution. Our evaluation on the University of Dundee 50 Salads, Georgia Tech Egocentric Activities, and JHU-ISI Gesture and Skill Assessment Working Set demonstrates that TDRN outperforms the state of the art in frame-wise segmentation accuracy, segmental edit score, and segmental overlap F1 score.
****

Weakly Supervised Action Localization by Sparse Temporal Pooling Network

Phuc Nguyen, Ting Liu, Gautam Prasad, Bohyung Han; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6752-6761

We propose a weakly supervised temporal action localization algorithm on untrimmed videos using convolutional neural networks. Our algorithm learns from video-level class labels and predicts temporal intervals of human actions with no requirement of temporal localization annotations. We design our network to identify a sparse subset of key segments associated with target actions in a video using an attention module and fuse the key segments through adaptive temporal pooling. Our loss function is comprised of two terms that minimize the video-level action classification error and enforce the sparsity of the segment selection. At inference time, we extract and score temporal proposals using temporal class activations and class-agnostic attentions to estimate the time intervals that correspond to target actions. The proposed algorithm attains state-of-the-art results on the THUMOS14 dataset and outstanding performance on ActivityNet1.3 even with its weak supervision.
****

PoseFlow: A Deep Motion Representation for Understanding Human Behaviors in Videos

Dingwen Zhang, Guangyu Guo, Dong Huang, Junwei Han; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6762-6770

Motion of the human body is the critical cue for understanding and characterizing human behavior in videos. Most existing approaches explore the motion cue using optical flows. However, optical flow usually contains motion on both the interested human bodies and the undesired background. This "noisy" motion representation makes it very challenging for pose estimation and action recognition in real scenarios. To address this issue, this paper presents a novel deep motion representation, called PoseFlow, which reveals human motion in videos while suppressing background and motion blur, and being robust to occlusion. For learning PoseFlow with mild computational cost, we propose a functionally structured spatial-temporal deep network, PoseFlow Net (PFN), to jointly solve the skeleton localization and matching problems of PoseFlow. Comprehensive experiments show that PFN outperforms the state-of-the-art deep flow estimation models in generating PoseFlow. Moreover, PoseFlow demonstrates its potential on improving two challenging tasks in human video analysis: pose estimation and action recognition.

****

FFNet: Video Fast-Forwarding via Reinforcement Learning

Shuyue Lan, Rameswar Panda, Qi Zhu, Amit K. Roy-Chowdhury; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6771-6780

For many intelligent applications with limited computation, communication, storage and energy resources, there is an imperative need of vision methods that could select an informative subset of the input video for efficient processing at or near real time. In the literature, there are two relevant groups of approaches: generating a "trailer" for a video or fast-forwarding while watching/processing the video. The first group is supported by video summarization techniques, which require processing of the entire video to select an important subset for showing to users. In the second group, current fast-forwarding methods depend on either manual control or automatic adaptation of playback speed, which often do not present an accurate representation and may still require processing of every frame. In this paper, we introduce FastForwardNet (FFNet), a reinforcement learning agent that gets inspiration from video summarization and does fast-forwarding differently. It is an online framework that automatically fast-forwards a video and presents a representative subset of frames to users on the fly. It does not require processing the entire video but just the portion that is selected by the fast-forward agent, which makes the process very computationally efficient. The online nature of our proposed method also enables the users to begin fast-forwarding at any point of the video. Experiments on two real-world datasets demonstrate that our method can provide better representation of the input video (about 6%-20% improvement on coverage of important frames) with much less processing requirement (more than 80% reduction in the number of frames processed).

****

Multi-Shot Pedestrian Re-Identification via Sequential Decision Making

Jianfu Zhang, Naiyan Wang, Liqing Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6781-6789

Multi-shot pedestrian re-identification problem is at the core of surveillance video analysis. It matches two tracks of pedestrians from different cameras. In contrary to existing works that aggregate single frames features by time series model such as recurrent neural network, in this paper, we propose an interpretable reinforcement learning based approach to this problem. Particularly, we train an agent to verify a pair of images at each time. The agent could choose to output the result (same or different) or request another pair of images to verify (unsure). By this way, our model implicitly learns the difficulty of image pairs, and postpone the decision when the model does not accumulate enough evidence. Moreover, by adjusting the reward for unsure action, we can easily trade off between speed and accuracy. In three open benchmarks, our method are competitive with the state-of-the-art methods while only using 3% to 6% images. These promising results demonstrate that our method is favorable in both efficiency and performance.

****

Attend and Interact: Higher-Order Object Interactions for Video Understanding

Chih-Yao Ma, Asim Kadav, Iain Melvin, Zsolt Kira, Ghassan AlRegib, Hans Peter Graf; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6790-6800

Human actions often involve complex interactions across several inter-related objects in the scene. However, existing approaches to fine-grained video understan ding or visual relationship detection often rely on single object representation or pairwise object relationships. Furthermore, learning interactions across multiple objects in hundreds of frames for video is computationally infeasible and performance may suffer since a large combinatorial space has to be modeled. In this paper, we propose to efficiently learn higher-order interactions between arbitrary subgroups of objects for fine-grained video understanding. We demonstrate that modeling object interactions significantly improves accuracy for both action recognition and video captioning, while saving more than 3-times the computation over traditional pairwise relationships. The proposed method is validated on two large-scale datasets: Kinetics and ActivityNet Captions. Our SINet and SINet-Caption achieve state-of-the-art performances on both datasets even though the videos are sampled at a maximum of 1 FPS. To the best of our knowledge, this is the first work modeling object interactions on open domain large-scale video datasets, and we additionally model higher-order object interactions which improves the performance with low computational costs.

****

Where and Why Are They Looking? Jointly Inferring Human Attention and Intentions in Complex Tasks

Ping Wei, Yang Liu, Tianmin Shu, Nanning Zheng, Song-Chun Zhu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6801-6809

This paper addresses a new problem - jointly inferring human attention, intentions, and tasks from videos. Given an RGB-D video where a human performs a task, we answer three questions simultaneously: 1) where the human is looking - attention prediction; 2) why the human is looking there - intention prediction; and 3) what task the human is performing - task recognition. We propose a hierarchical model of human-attention-object (HAO) which represents tasks, intentions, and attention under a unified framework. A task is represented as sequential intentions which transition to each other. An intention is composed of the human pose, attention, and objects. A beam search algorithm is adopted for inference on the HAO graph to output the attention, intention, and task results. We built a new video dataset of tasks, intentions, and attention. It contains 14 task classes, 70 intention categories, 28 object classes, 809 videos, and approximately 330,000 frames. Experiments show that our approach outperforms existing approaches.

****

Fully Convolutional Adaptation Networks for Semantic Segmentation

Yiheng Zhang, Zhaofan Qiu, Ting Yao, Dong Liu, Tao Mei; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6810-6818

The recent advances in deep neural networks have convincingly demonstrated high capability in learning vision models on large datasets. Nevertheless, collecting expert labeled datasets especially with pixel-level annotations is an extremely expensive process. An appealing alternative is to render synthetic data (e.g., computer games) and generate ground truth automatically. However, simply applying the models learnt on synthetic images may lead to high generalization error on real images due to domain shift. In this paper, we facilitate this issue from the perspectives of both visual appearance-level and representation-level domain adaptation. The former adapts source-domain images to appear as if drawn from the ``style" in the target domain and the latter attempts to learn domain-invariant representations. Specifically, we present Fully Convolutional Adaptation Networks (FCAN), a novel deep architecture for semantic segmentation which combines Appearance Adaptation Networks (AAN) and Representation Adaptation Networks (RAN). AAN learns a transformation from one domain to the other in the pixel space and RAN is optimized in an adversarial learning manner to maximally fool the domain discriminator with the learnt source and target representations. Extensive experiments are conducted on the transfer from GTA5 (game videos) to Cityscapes (urban street scenes) on semantic segmentation and our proposal achieves superior results when comparing to state-of-the-art unsupervised adaptation techniques. More remarkably, we obtain a new record: mIoU of 47.5% on BDDS (drive-cam videos) in an unsupervised setting.

****

Semantic Video Segmentation by Gated Recurrent Flow Propagation

David Nilsson, Cristian Sminchisescu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6819-6828

Semantic video segmentation is challenging due to the sheer amount of data that needs to be processed and labeled in order to construct accurate models. In this paper we present a deep, end-to-end trainable methodology for video segmentation that is capable of leveraging the information present in unlabeled data, besides sparsely labeled frames, in order to improve semantic estimates. Our model combines a convolutional architecture and a spatio-temporal transformer recurrent layer that is able to temporally propagate labeling information by means of optical flow, adaptively gated based on its locally estimated uncertainty. The flow, the recognition and the gated temporal propagation modules can be trained jointly, end-to-end. The temporal, gated recurrent flow propagation component of our model can be plugged into any static semantic segmentation architecture and turn it into a weakly supervised video processing one. Our experiments in the challenging CityScapes and Camvid datasets, and for multiple deep architectures, indicate that the resulting model can leverage unlabeled temporal frames, next to a labeled one, in order to improve both the video segmentation accuracy and the consistency of its temporal labeling, at no additional annotation cost and with little extra computation.

****

Interpretable Video Captioning via Trajectory Structured Localization

Xian Wu, Guanbin Li, Qingxing Cao, Qingge Ji, Liang Lin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6829-6837

Automatically describing open-domain videos with natural language are attracting increasing interest in the field of artificial intelligence. Most existing methods simply borrow ideas from image captioning and obtain a compact video representation from an ensemble of global image feature before feeding to an RNN decoder which outputs a sentence of variable length. However, it is not only arduous for the generator to focus on specific salient objects at different time given the global video representation, it is more formidable to capture the fine-grained motion information and the relation between moving instances for more subtle linguistic descriptions. In this paper, we propose a Trajectory Structured Attentional Encoder-Decoder (TSA-ED) neural network framework for more elaborate video captioning which works by integrating local spatial-temporal representation at trajectory level through structured attention mechanism. Our proposed method is based on a LSTM-based encoder-decoder framework, which incorporates an attention modeling scheme to adaptively learn the correlation between sentence structure and the moving objects in videos, and consequently generates more accurate and meticulous statement description in the decoding stage. Experimental results demonstrate that the feature representation and structured attention mechanism based on the trajectory cluster can efficiently obtain the local motion information in the video to help generate a more fine-grained video description, and achieve the state-of-the-art performance on the well-known Charades and MSVD datasets.

****

Deep Hashing via Discrepancy Minimization

Zhixiang Chen, Xin Yuan, Jiwen Lu, Qi Tian, Jie Zhou; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6838-6847

This paper presents a discrepancy minimizing model to address the discrete optimization problem in hashing learning. The discrete optimization introduced by binary constraint is an NP-hard mixed integer programming problem. It is usually addressed by relaxing the binary variables into continuous variables to adapt to the gradient based learning of hashing functions, especially the training of deep neural networks. To deal with the objective discrepancy caused by relaxation, we transform the original binary optimization into differentiable optimization problem over hash functions through series expansion. This transformation decouples the binary constraint and the similarity preserving hashing function optimization. The transformed objective is optimized in a tractable alternating optimiza tion framework with gradual discrepancy minimization. Extensive experimental results on three benchmark datasets validate the efficacy of the proposed discrepancy minimizing hashing.

****

ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices

Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, Jian Sun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6848-6856

We introduce an extremely computation-efficient CNN architecture named ShuffleNet, which is designed specially for mobile devices with very limited computing power (e.g., 10-150 MFLOPs). The new architecture utilizes two new operations, pointwise group convolution and channel shuffle, to greatly reduce computation cost while maintaining accuracy. Experiments on ImageNet classification and MS COCO object detection demonstrate the superior performance of ShuffleNet over other structures, e.g. lower top-1 error (absolute 7.8%) than recent MobileNet~cite{howard2017mobilenets} on ImageNet classification task, under the computation budget of 40 MFLOPs. On an ARM-based mobile device, ShuffleNet achieves $sim$13$■imes$ actual speedup over AlexNet while maintaining comparable accuracy.

****

Zero-Shot Recognition via Semantic Embeddings and Knowledge Graphs

Xiaolong Wang, Yufei Ye, Abhinav Gupta; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6857-6866

We consider the problem of zero-shot recognition: learning a visual classifier for a category with zero training examples, just using the word embedding of the category and its relationship to other categories, which visual data are provided. The key to dealing with the unfamiliar or novel category is to transfer knowledge obtained from familiar classes to describe the unfamiliar class. In this paper, we build upon the recently introduced Graph Convolutional Network (GCN) and propose an approach that uses both semantic embeddings and the categorical relationships to predict the classifiers. Given a learned knowledge graph (KG), our approach takes as input semantic embeddings for each node (representing visual category). After a series of graph convolutions, we predict the visual classifier for each category. During training, the visual classifiers for a few categories are given to learn the GCN parameters. At test time, these filters are used to predict the visual classifiers of unseen categories. We show that our approach is robust to noise in the KG. More importantly, our approach provides significant improvement in performance compared to the current state-of-the-art results (from 2 ~ 3% on some metrics to whopping 20% on a few).

****

Referring Relationships

Ranjay Krishna, Ines Chami, Michael Bernstein, Li Fei-Fei; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6867-6876

Images are not simply sets of objects: each image represents a web of interconnected relationships. These relationships between entities carry semantic meaning and help a viewer differentiate between instances of an entity. For example, in an image of a soccer match, there may be multiple persons present, but each participates in different relationships: one is kicking the ball, and the other is guarding the goal. In this paper, we formulate the task of utilizing these "referring relationships" to disambiguate between entities of the same category. We introduce an iterative model that localizes the two entities in the referring relationship, conditioned on one another. We formulate the cyclic condition between the entities in a relationship by modelling predicates that connect the entities as shifts in attention from one entity to another. We demonstrate that our model can not only outperform existing approaches on three datasets --- CLEVR, VRD and Visual Genome --- but also that it produces visually meaningful predicate shifts, as an instance of interpretable neural networks. Finally, we show that by modelling predicates as attention shifts, we can even localize entities in the absence of their category, allowing our model to find completely unseen categories.

****

Improving Object Localization With Fitness NMS and Bounded IoU Loss

Lachlan Tychsen-Smith, Lars Petersson; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6877-6885

We demonstrate that many detection methods are designed to identify only a sufficiently accurate bounding box, rather than the best available one. To address this issue we propose a simple and fast modification to the existing methods called Fitness NMS. This method is tested with the DeNet model and obtains a significantly improved MAP at greater localization accuracies without a loss in evaluation rate, and can be used in conjunction with Soft NMS for additional improvements. Next we derive a novel bounding box regression loss based on a set of IoU upper bounds that better matches the goal of IoU maximization while still providing good convergence properties. Following these novelties we investigate RoI clustering schemes for improving evaluation rates for the DeNet wide model variants and provide an analysis of localization performance at various input image dimensions. We obtain a MAP of 33.6%@79Hz and 41.8%@5Hz for MSCOCO and a Titan X (Maxwell).

****

End-to-End Deep Kronecker-Product Matching for Person Re-Identification

Yantao Shen, Tong Xiao, Hongsheng Li, Shuai Yi, Xiaogang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6886-6895

Person re-identification aims to robustly measure similarities between person images. The significant variation of person poses and viewing angles challenges for accurate person re-identification. The spatial layout and correspondences between query person images are vital information for tackling this problem but are ignored by most state-of-the-art methods. In this paper, we propose a novel Kronecker Product Matching module to match feature maps of different persons in an end-to-end trainable deep neural network. A novel feature soft warping scheme is designed for aligning the feature maps based on matching results, which is shown to be crucial for achieving superior accuracy. The multi-scale features based on hourglass-like networks and self residual attention are also exploited to further boost the re-identification performance. The proposed approach outperforms state-of-the-art methods on the Market-1501, CUHK03, and DukeMTMC datasets, which demonstrates the effectiveness and generalization ability of our proposed approach.

****

Semantic Visual Localization

Johannes L. Schönberger, Marc Pollefeys, Andreas Geiger, Torsten Sattler; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6896-6906

Robust visual localization under a wide range of viewing conditions is a fundamental problem in computer vision. Handling the difficult cases of this problem is not only very challenging but also of high practical relevance, e.g., in the context of life-long localization for augmented reality or autonomous robots. In this paper, we propose a novel approach based on a joint 3D geometric and semantic understanding of the world, enabling it to succeed under conditions where previous approaches failed. Our method leverages a novel generative model for descriptor learning, trained on semantic scene completion as an auxiliary task. The resulting 3D descriptors are robust to missing observations by encoding high-level 3D geometric and semantic information. Experiments on several challenging large-scale localization datasets demonstrate reliable localization under extreme viewpoint, illumination, and geometry changes.

****

Objects as Context for Detecting Their Semantic Parts

Abel Gonzalez-Garcia, Davide Modolo, Vittorio Ferrari; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6907-6916

We present a semantic part detection approach that effectively leverages object information. We use the object appearance and its class as indicators of what parts to expect. We also model the expected relative location of parts inside the objects based on their appearance. We achieve this with a new network module, called OffsetNet, that efficiently predicts a variable number of part locations within a given object. Our model incorporates all these cues to detect parts in the context of their objects. This leads to considerably higher performance for the challenging task of part detection compared to using part appearance alone (+5 mAP on the PASCAL-Part dataset). We also compare to other part detection methods on both PASCAL-Part and CUB200-2011 datasets.

****

End-to-End Weakly-Supervised Semantic Alignment

Ignacio Rocco, Relja Arandjelovi■, Josef Sivic; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6917-6925

We tackle the task of semantic alignment where the goal is to compute dense semantic correspondence aligning two images depicting objects of the same category. This is a challenging task due to large intra-class variation, changes in viewpoint and background clutter.   We present the following three principal contributions. First, we develop a convolutional neural network architecture for semantic alignment that is trainable in an end-to-end manner from weak image-level supervision in the form of matching image pairs. The outcome is that parameters are learnt from rich appearance variation present in different but semantically related images without the need for tedious manual annotation of correspondences at training time. Second, the main component of this architecture is a differentiable soft inlier scoring module, inspired by the RANSAC inlier scoring procedure, that computes the quality of the alignment based on only geometrically consistent correspondences thereby reducing the effect of background clutter. Third, we demonstrate that the proposed approach achieves state-of-the-art performance on multiple standard benchmarks for semantic alignment.

****

Dynamic Zoom-In Network for Fast Object Detection in Large Images

Mingfei Gao, Ruichi Yu, Ang Li, Vlad I. Morariu, Larry S. Davis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6926-6935

We introduce a generic framework that reduces the computational cost of object detection while retaining accuracy for scenarios where objects with varied sizes appear in high resolution images. Detection progresses in a coarse-to-fine manner, first on a down-sampled version of the image and then on a sequence of higher resolution regions identified as likely to improve the detection accuracy. Built upon reinforcement learning, our approach consists of a model (R-net) that uses coarse detection results to predict the potential accuracy gain for analyzing a region at a higher resolution and another model (Q-net) that sequentially selects regions to zoom in. Experiments on the Caltech Pedestrians dataset show that our approach reduces the number of processed pixels by over 50% without a drop in detection accuracy. The merits of our approach become more significant on a high resolution test set collected from YFCC100M dataset, where our approach maintains high detection performance while reducing the number of processed pixels by about 70% and the detection time by over 50%.

****

Learning Markov Clustering Networks for Scene Text Detection

Zichuan Liu, Guosheng Lin, Sheng Yang, Jiashi Feng, Weisi Lin, Wang Ling Goh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6936-6944

A novel framework named Markov Clustering Network (MCN) is proposed for fast and robust scene text detection. MCN predicts instance-level bounding boxes by firstly converting an image into a Stochastic Flow Graph (SFG) and then performing Markov Clustering on this graph. Our method can detect text objects with arbitrary size and orientation without prior knowledge of object size. The stochastic flow graph encode objects' local correlation and semantic information. An object is modeled as strongly connected nodes, which allows flexible bottom-up detection for scale-varying and rotated objects. MCN generates bounding boxes without using Non-Maximum Suppression, and it can be fully parallelized on GPUs. The evaluation on public benchmarks shows that our method outperforms the existing methods by a large margin in detecting multioriented text objects. MCN achieves new state-of-art performance on challenging MSRA-TD500 dataset with precision of 0.88, recall of 0.79 and F-score of 0.83. Also, MCN achieves realtime inference with frame rate of 34 FPS, which is $1.5■imes$ speedup when compared with the fastest scene text detection algorithm.

****

Deep Reinforcement Learning of Region Proposal Networks for Object Detection

Aleksis Pirinen, Cristian Sminchisescu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6945-6954

We propose drl-RPN, a deep reinforcement learning-based visual recognition model consisting of a sequential region proposal network (RPN) and an object detector. In contrast to typical RPNs, where candidate object regions (RoIs) are selected greedily via class-agnostic NMS, drl-RPN optimizes an objective closer to the final detection task. This is achieved by replacing the greedy RoI selection process with a sequential attention mechanism which is trained via deep reinforcement learning (RL). Our model is capable of accumulating class-specific evidence over time, potentially affecting subsequent proposals and classification scores, and we show that such context integration significantly boosts detection accuracy. Moreover, drl-RPN automatically decides when to stop the search process and has the benefit of being able to jointly learn the parameters of the policy and the detector, both represented as deep networks. Our model can further learn to search over a wide range of exploration-accuracy trade-offs making it possible to specify or adapt the exploration extent at test time. The resulting search trajectories are image- and category-dependent, yet rely only on a single policy over all object categories. Results on the MS COCO and PASCAL VOC challenges show that our approach outperforms established, typical state-of-the-art object detection pipelines.

****

Beyond Holistic Object Recognition: Enriching Image Understanding With Part States

Cewu Lu, Hao Su, Yonglu Li, Yongyi Lu, Li Yi, Chi-Keung Tang, Leonidas J. Guibas; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6955-6963

Important high-level vision tasks require rich semantic descriptions of objects at part level. Based upon previous work on part localization, in this paper, we address the problem of inferring rich semantics imparted by an object part in still images. Specifically, we propose to tokenize the semantic space as a discrete set of part states. Our modeling of part state is spatially localized, therefore, we formulate the part state inference problem as a pixel-wise annotation problem. An iterative part-state inference neural network that is efficient in time and accurate in performance is specifically designed for this task. Extensive experiments demonstrate that the proposed method can effectively predict the semantic states of parts and simultaneously improve part segmentation, thus benefiting a number of visual understanding applications. The other contribution of this paper is our part state dataset which contains rich part-level semantic annotations.

****

Discriminability Objective for Training Descriptive Captions

Ruotian Luo, Brian Price, Scott Cohen, Gregory Shakhnarovich; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6964-6974

One property that remains lacking in image captions generated by contemporary methods is discriminability: being able to tell two images apart given the caption for one of them. We propose a way to improve this aspect of caption generation. By incorporating into the captioning training objective a loss component directly related to ability (by a machine) to disambiguate image/caption matches, we obtain systems that produce much more discriminative caption, according to human evaluation. Remarkably, our approach leads to improvement in other aspects of generated captions, reflected by a battery of standard scores such as BLEU, SPICE etc. Our approach is modular and can be applied to a variety of model/loss combi nations commonly proposed for image captioning.

****

Visual Question Answering With Memory-Augmented Networks

Chao Ma, Chunhua Shen, Anthony Dick, Qi Wu, Peng Wang, Anton van den Hengel, Ian Reid; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6975-6984

In this paper, we exploit memory-augmented neural networks to predict accurate answers to visual questions, even when those answers rarely occur in the training set. The memory network incorporates both internal and external memory blocks and selectively pays attention to each training exemplar. We show that memory-augmented neural networks are able to maintain a relatively long-term memory of scarce training exemplars, which is important for visual question answering due to the heavy-tailed distribution of answers in a general VQA setting. Experimental results in two large-scale benchmark datasets show the favorable performance of the proposed algorithm with the comparison to state of the art.

****

Structure Inference Net: Object Detection Using Scene-Level Context and Instance-Level Relationships

Yong Liu, Ruiping Wang, Shiguang Shan, Xilin Chen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6985-6994

Context is important for accurate visual recognition. In this work we propose an object detection algorithm that not only considers object visual appearance, but also makes use of two kinds of context including scene contextual information and object relationships within a single image. Therefore, object detection is regarded as both a cognition problem and a reasoning problem when leveraging these structured information. Specifically, this paper formulates object detection as a problem of graph structure inference, where given an image the objects are treated as nodes in a graph and relationships between the objects are modeled as edges in such graph. To this end, we present a so-called Structure Inference Network (SIN), a detector that incorporates into a typical detection framework (e.g. Faster R-CNN) with a graphical model which aims to infer object state. Comprehensive experiments on PASCAL VOC and MS COCO datasets indicate that scene context and object relationships truly improve the performance of object detection with more desirable and reasonable outputs.

****

Occluded Pedestrian Detection Through Guided Attention in CNNs

Shanshan Zhang, Jian Yang, Bernt Schiele; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6995-7003

Pedestrian detection has progressed significantly in the last years. However, occluded people are notoriously hard to detect, as their appearance varies substantially depending on a wide range of partial occlusions. In this paper, we aim to propose a simple and compact method based on the FasterRCNN architecture for occluded pedestrian detection. We start with interpreting CNN channel features of a pedestrian detector, and we find that different channels activate responses for different body parts respectively. These findings strongly motivate us to employ an attention mechanism across channels to represent various occlusion patterns in one single model, as each occlusion pattern can be formulated as some specific combination of body parts. Therefore, an attention network with self or external guidances is proposed as an add-on to the baseline FasterRCNN detector. When evaluating on the heavy occlusion subset, we achieve a significant improvement of 8pp to the baseline FasterRCNN detector on CityPersons and on Caltech we outperform the state-of-the-art method by 4pp.

****

Reward Learning From Narrated Demonstrations

Hsiao-Yu Tung, Adam W. Harley, Liang-Kang Huang, Katerina Fragkiadaki; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7004-7013

Humans effortlessly “program” one another by communicating goals and desires in natural language. In contrast, humans program robotic behaviours by indicating desired object locations and poses to be achieved [5], by providing RGB images of goal configurations [19], or supplying a demonstration to be imitated [17]. None of these methods generalize across environment variations, and they convey the goal in awkward technical terms. This work proposes joint learning of natural language grounding and instructable behavioural policies reinforced by perceptual detectors of natural language expressions, grounded to the sensory inputs of the robotic agent. Our supervision is narrated visual demonstrations (NVD), which are visual demonstrations paired with verbal narration (as opposed to being silent). We introduce a dataset of NVD where teachers perform activities while describing them in detail. We map the teachers’ descriptions to perceptual reward detectors, and use them to train corresponding behavioural policies in simulation. We empirically show that our instructable agents (i) learn visual reward detectors using a small number of examples by exploiting hard negative mined configurations from demonstration dynamics, (ii) develop pick-and-place policies using learned visual reward detectors, (iii) benefit from object-factorized state representations that mimic the syntactic structure of natural language goal expressions, and (iv) can execute behaviours that involve novel objects in novel locations at test time, instructed by natural language.

****

Weakly-Supervised Semantic Segmentation Network With Deep Seeded Region Growing

Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, Jingdong Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7014-7023

This paper studies the problem of learning image semantic segmentation networks only using image-level labels as supervision, which is important since it can significantly reduce human annotation efforts. Recent state-of-the-art methods on this problem first infer the sparse and discriminative regions for each object class using a deep classification network, then train semantic a segmentation network using the discriminative regions as supervision. Inspired by the traditional image segmentation methods of seeded region growing, we propose to train a semantic segmentation network starting from the discriminative regions and progressively increase the pixel-level supervision using by seeded region growing. The seeded region growing module is integrated in a deep segmentation network and can benefit from deep features. Different from conventional deep networks which have fixed/static labels, the proposed weakly-supervised network generates new labels using the contextual information within an image. The proposed method significantly outperforms the weakly-supervised semantic segmentation methods using static labels, and obtains the state-of-the-art performance, which are 63.2% mIoU score on the PASCAL VOC 2012 test set and 26.0% mIoU score on the COCO dataset.

****

PoTion: Pose MoTion Representation for Action Recognition

Vasileios Choutas, Philippe Weinzaepfel, Jérôme Revaud, Cordelia Schmid; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7024-7033

Most state-of-the-art methods for action recognition rely on a two-stream architecture that processes appearance and motion independently. In this paper, we claim that considering them jointly offers rich information for action recognition. We introduce a novel representation that gracefully encodes the movement of some semantic keypoints. We use the human joints as these keypoints and term our Pose moTion representation PoTion. Specifically, we first run a state-of-the-art human pose estimator and extract heatmaps for the human joints in each frame. We obtain our PoTion representation by temporally aggregating these probability maps. This is achieved by colorizing each of them depending on the relative time of the frames in the video clip and summing them. This fixed-size representation for an entire video clip is suitable to classify actions using a shallow convolutional neural network. Our experimental evaluation shows that PoTion outperforms other state-of-the-art pose representations. Furthermore, it is complementary to standard appearance and motion streams. When combining PoTion with the recent two-stream I3D approach [5], we obtain state-of-the-art performance on the JHMDB, HMDB and UCF101 datasets.

****

Bilateral Ordinal Relevance Multi-Instance Regression for Facial Action Unit Intensity Estimation
Yong Zhang, Rui Zhao, Weiming Dong, Bao-Gang Hu, Qiang Ji; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7034-7043

Automatic intensity estimation of facial action units (AUs) is challenging in two aspects. First, capturing subtle changes of facial appearance is quiet difficult. Second, the annotation of AU intensity is scarce and expensive. Intensity annotation requires strong domain knowledge thus only experts are qualified. The majority of methods directly apply supervised learning techniques to AU intensity estimation while few methods exploit unlabeled samples to improve the performance. In this paper, we propose a novel weakly supervised regression model-Bilateral Ordinal Relevance Multi-instance Regression (BORMIR), which learns a frame-level intensity estimator with weakly labeled sequences. From a new perspective, we introduce relevance to model sequential data and consider two bag labels for each bag. The AU intensity estimation is formulated as a joint regressor and relevance learning problem. Temporal dynamics of both relevance and AU intensity are leveraged to build connections among labeled and unlabeled image frames to provide weak supervision. We also develop an efficient algorithm for optimization based on the alternating minimization framework. Evaluations on three expression databases demonstrate the effectiveness of the proposed model.

****

Pulling Actions out of Context: Explicit Separation for Effective Combination
Yang Wang, Minh Hoai; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7044-7053

The ability to recognize human actions in video has many potential applications. Human action recognition, however, is tremendously challenging for computers due to the complexity of video data and the subtlety of human actions. Most current recognition systems flounder on the inability to separate human actions from co-occurring factors that usually dominate subtle human actions. In this paper, we propose a novel approach for training a human action recognizer, one that can: (1) explicitly factorize human actions from the co-occurring factors; (2) deliberately build a model for human actions and a separate model for all correlated contextual elements; and (3) effectively combine the models for human action recognition. Our approach exploits the benefits of conjugate samples of human actions, which are video clips that are contextually similar to human action samples, but do not contain the action. Experiments on ActionThread, PASCAL VOC, UCF101, and Hollywood2 datasets demonstrate the ability to separate action from context of the proposed approach.

****

Dynamic Feature Learning for Partial Face Recognition
Lingxiao He, Haiqing Li, Qi Zhang, Zhenan Sun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7054-7063

Partial face recognition (PFR) in unconstrained environment is a very important task, especially in video surveillance, mobile devices, etc. However, a few studies have tackled how to recognize an arbitrary patch of a face image. This study combines Fully Convolutional Network (FCN) with Sparse Representation Classification (SRC) to propose a novel partial face recognition approach, called Dynamic Feature Matching (DFM), to address partial face images regardless of sizes. Based on DFM, we propose a sliding loss to optimize FCN by reducing the intra-variation between a face patch and face images of a subject, which further improves the performance of DFM. The proposed DFM is evaluated on several partial face databases, including LFW, YTF and CASIA-NIR-Distance databases. Experimental results demonstrate the effectiveness and advantages of DFM in comparison with state-of-the-art PFR methods.

****

Exploiting Transitivity for Learning Person Re-Identification Models on a Budget
Sourya Roy, Sujoy Paul, Neal E. Young, Amit K. Roy-Chowdhury; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7064-7072

Minimization of labeling effort for person re-identification in camera networks is an important problem as most of the existing popular methods are supervised and they require large amount of manual annotations, acquiring which is a tedious job. In this work, we focus on this labeling effort minimization problem and approach it as a subset selection task where the objective is to select an optimal subset of image-pairs for labeling without compromising performance. Towards this goal, our proposed scheme first represents any camera network (with k number of cameras) as an edge weighted complete k-partite graph where each vertex denotes a person and similarity scores between persons are used as edge-weights. Then in the second stage, our algorithm selects an optimal subset of pairs by solving a triangle free subgraph maximization problem on the k-partite graph. This sub-graph weight maximization problem is NP-hard (at least for k > = 4) which means for large datasets the optimization problem becomes intractable. In order to make our framework scalable, we propose two polynomial time approximately-optimal algorithms. The first algorithm is a 1/2-approximation algorithm which runs in linear time in the number of edges. The second algorithm is a greedy algorithm with sub-quadratic (in number of edges) time-complexity. Experiments on three state-of-the-art datasets depict that the proposed approach requires on an average only 8-15 % manually labeled pairs in order to achieve the performance when all the pairs are manually annotated.
****

Deep Spatial Feature Reconstruction for Partial Person Re-Identification: Alignment-Free Approach

Lingxiao He, Jian Liang, Haiqing Li, Zhenan Sun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7073-7082

Partial person re-identification (re-id) is a challenging problem, where only a partial observation of a person image is available for matching. However, few studies have offered a solution of how to identify an arbitrary patch of a person image. In this paper, we propose a fast and accurate matching method to address this problem. The proposed method leverages Fully Convolutional Network (FCN) to generate correspondingly-size spatial feature maps such that pixel-level features are consistent. To match a pair of person images of different sizes, a novel method called Deep Spatial feature Reconstruction (DSR) is further developed to avoid explicit alignment. Specifically, we exploit the reconstructing error from dictionary learning to calculate the similarity between different spatial feature maps. In that way, we expect that the proposed FCN can decrease the similarity of coupled images from different persons and vice versa. Experimental results on two partial person datasets demonstrate the efficiency and effectiveness of the proposed method in comparison with several state-of-the-art partial person re-id approaches.
****

Every Smile Is Unique: Landmark-Guided Diverse Smile Generation

Wei Wang, Xavier Alameda-Pineda, Dan Xu, Pascal Fua, Elisa Ricci, Nicu Sebe; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7083-7092

Each smile is unique: one person surely smiles in different ways (e.g., closing/opening the eyes or mouth). Given one input image of a neutral face, can we generate multiple smile videos with distinctive characteristics? To tackle this one-to-many video generation problem, we propose a novel deep learning architecture named Conditional Multi-Mode Network (CMM-Net). To better encode the dynamics of facial expressions, CMM-Net explicitly exploits facial landmarks for generating smile sequences. Specifically, a variational auto-encoder is used to learn a facial landmark embedding. This single embedding is then exploited by a conditional recurrent network which generates a landmark embedding sequence conditioned on a specific expression (e.g., spontaneous smile). Next, the generated landmark embeddings are fed into a multi-mode recurrent landmark generator, producing a set of landmark sequences still associated to the given smile class but clearly distinct from each other. Finally, these landmark sequences are translated into face videos. Our experimental results demonstrate the effectiveness of our CMM-Net in generating realistic videos of multiple smile expressions.

****

UV-GAN: Adversarial Facial UV Map Completion for Pose-Invariant Face Recognition

Jiankang Deng, Shiyang Cheng, Niannan Xue, Yuxiang Zhou, Stefanos Zafeiriou; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7093-7102

Recently proposed robust 3D face alignment methods establish either dense or sparse correspondence between a 3D face model and a 2D facial image. The use of these methods presents new challenges as well as opportunities for facial texture analysis. In particular, by sampling the image using the fitted model, a facial UV can be created. Unfortunately, due to self-occlusion, such a UV map is always incomplete. In this paper, we propose a framework for training Deep Convolutional Neural Network (DCNN) to complete the facial UV map extracted from in-the-wild images. To this end, we first gather complete UV maps by fitting a 3D Morphable Model (3DMM) to various multiview image and video datasets, as well as leveraging on a new 3D dataset with over 3,000 identities. Second, we devise a meticulously designed architecture that combines local and global adversarial DCNNs to learn an identity-preserving facial UV completion model. We demonstrate that by attaching the completed UV to the fitted mesh and generating instances of arbitrary poses, we can increase pose variations for training deep face recognition/verification models, and minimise pose discrepancy during testing, which lead to better performance. Experiments on both controlled and in-the-wild UV datasets prove the effectiveness of our adversarial UV completion model. We achieve state-of-the-art verification accuracy, 94.05%, under the CFP frontal-profile protocol only by combining pose augmentation during training and pose discrepancy reduction during testing. We will release the first in-the-wild UV dataset (we refer as WildUV) that comprises of complete facial UV maps from 1,892 identities for research purposes.

****

Cascaded Pyramid Network for Multi-Person Pose Estimation

Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, Jian Sun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7103-7112

The topic of multi-person pose estimation has beenlargely improved recently, especially with the developmentof convolutional neural network. However, there still exista lot of challenging cases, such as occluded keypoints, in-visible keypoints and complex background, which cannot bewell addressed. In this paper, we present a novel networkstructure called Cascaded Pyramid Network (CPN) whichtargets to relieve the problem from these “hard” keypoints.More specifically, our algorithm includes two stages: Glob-alNet and RefineNet. GlobalNet is a feature pyramid net-work which can successfully localize the “simple” key-points like eyes and hands but may fail to precisely rec-ognize the occluded or invisible keypoints. Our RefineNettries explicitly handling the “hard” keypoints by integrat-ing all levels of feature representations from the Global-Net together with an online hard keypoint mining loss. Ingeneral, to address the multi-person pose estimation prob-lem, a top-down pipeline is adopted to first generate a setof human bounding boxes based on a detector, followed byour CPN for keypoint localization in each human boundingbox. Based on the proposed algorithm, we achieve state-of-art results on the COCO keypoint benchmark, with averageprecision at 73.0 on the COCO test-dev dataset and 72.1 onthe COCO test-challenge dataset, which is a 19% relativeimprovement compared with 60.5 from the COCO 2016 key-point challenge. Code and the detection results for personused will be publicly available for further research.

****

A Face-to-Face Neural Conversation Model

Hang Chu, Daiqing Li, Sanja Fidler; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7113-7121

Neural networks have recently become good at engaging in dialog. However, current approaches are based solely on verbal text, lacking the richness of a real face-to-face conversation. We propose a neural conversation model that aims to read and generate facial gestures alongside with text. This allows our model to adap t its response based on the “mood” of the conversation. In particular, we introduce an RNN encoder-decoder that exploits the movement of facial muscles, as well as the verbal conversation. The decoder consists of two layers, where the lower layer aims at generating the verbal response and coarse facial expressions, while the second layer fills in the subtle gestures, making the generated output more smooth and natural. We train our neural network by having it “watch” 250 movies. We showcase our joint face-text model in generating more natural conversations through automatic metrics and a human study. We demonstrate an example application with a face-to-face chatting avatar.

****

End-to-End Recovery of Human Shape and Pose

Angjoo Kanazawa, Michael J. Black, David W. Jacobs, Jitendra Malik; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7122-7131

We describe Human Mesh Recovery (HMR), an end-to-end framework for reconstructing a full 3D mesh of a human body from a single RGB image. In contrast to most current methods that compute 2D or 3D joint locations, we produce a richer and more useful mesh representation that is parameterized by shape and 3D joint angles. The main objective is to minimize the reprojection loss of keypoints, which allows our model to be trained using in-the-wild images that only have ground truth 2D annotations. However, the reprojection loss alone is highly underconstrained. In this work we address this problem by introducing an adversary trained to tell whether human body shape and pose are real or not using a large database of 3D human meshes. We show that HMR can be trained with and without using any paired 2D-to-3D supervision. We do not rely on intermediate 2D keypoint detections and infer 3D pose and shape parameters directly from image pixels. Our model runs in real-time given a bounding box containing the person. We demonstrate our approach on various images in-the-wild and out-perform previous optimization-based methods that output 3D meshes and show competitive results on tasks such as 3D joint location estimation and part segmentation.

****