Squeeze-and-Excitation Networks

Jie Hu, Li Shen, Gang Sun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7132-7141

Convolutional neural networks are built upon the convolution operation, which extracts informative features by fusing spatial and channel-wise information together within local receptive fields. In order to boost the representational power of a network, several recent approaches have shown the benefit of enhancing spatial encoding. In this work, we focus on the channel relationship and propose a novel architectural unit, which we term the “Squeeze-and-Excitation” (SE) block, that adaptively recalibrates channel-wise feature responses by explicitly modelling interdependencies between channels. We demonstrate that by stacking these blocks together, we can construct SENet architectures that generalise extremely well across challenging datasets. Crucially, we find that SE blocks produce significant performance improvements for existing state-of-the-art deep architectures at minimal additional computational cost. SENets formed the foundation of our ILSVRC 2017 classification submission which won first place and significantly reduced the top-5 error to 2.251%, achieving a ~25% relative improvement over the winning entry of 2016. Code and models are available at https: //github.com/hujie-frank/SENet.

****

Revisiting Salient Object Detection: Simultaneous Detection, Ranking, and Subitizing of Multiple Salient Objects

Md Amirul Islam, Mahmoud Kalash, Neil D. B. Bruce; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7142-7150

Salient object detection is a problem that has been considered in detail and many solutions proposed. In this paper, we argue that work to date has addressed a problem that is relatively ill-posed. Specifically, there is not universal agreement about what constitutes a salient object when multiple observers are queried. This implies that some objects are more likely to be judged salient than others, and implies a relative rank exists on salient objects. The solution presented in this paper solves this more general problem that considers relative rank, and we propose data and metrics suitable to measuring success in a relative object saliency landscape. A novel deep learning solution is proposed based on a hierarchical representation of relative saliency and stage-wise refinement. We also show that the problem of salient object subitizing can be addressed with the same network, and our approach exceeds performance of any prior work across all metrics considered (both traditional and newly proposed).

****

Context Encoding for Semantic Segmentation

Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, Amit Agrawal; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7151-7160

Recent work has made significant progress in improving spatial resolution for pixelwise labeling with Fully Convolutional Network (FCN) framework by employing Dilated/Atrous convolution, utilizing multi-scale features and refining boundaries. In this paper, we explore the impact of global contextual information in semantic segmentation by introducing the Context Encoding Module, which captures the semantic context of scenes and selectively highlights class-dependent featuremaps. The proposed Context Encoding Module significantly improves semantic segmentation results with only marginal extra computation cost over FCN. Our approach has achieved new state-of-the-art results 51.7% mIoU on PASCAL-Context, 85.9% mIoU on PASCAL VOC 2012. Our single model achieves a final score of 0.5567 on ADE20K test set, which surpass the winning entry of COCO-Place Challenge in 2017. In addition, we also explore how the Context Encoding Module can improve the feature representation of relatively shallow networks for the image classification on CIFAR-10 dataset. Our 14 layer network has achieved an error rate of 3.45%, which is comparable with state-of-the-art approaches with over 10 times more layers. The source code for the complete system are publicly available.

****

Creating Capsule Wardrobes From Fashion Images

Wei-Lin Hsiao, Kristen Grauman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7161-7170

We propose to automatically create emph{capsule wardrobes}. Given an inventory of candidate garments and accessories, the algorithm must assemble a minimal set of items that provides maximal mix-and-match outfits. We pose the task as a subset selection problem. To permit efficient subset selection over the space of all outfit combinations, we develop submodular objective functions capturing the key ingredients of visual compatibility, versatility, and user-specific preference. Since adding garments to a capsule only expands its possible outfits, we devise an iterative approach to allow near-optimal submodular function maximization. Finally, we present an unsupervised approach to learn visual compatibility from ``in the wild" full body outfit photos; the compatibility metric translates well to cleaner catalog photos and improves over existing methods. Our results on thousands of pieces from popular fashion websites show that automatic capsule creation has potential to mimic skilled fashionistas in assembling flexible wardrobes, while being significantly more scalable.

****

Webly Supervised Learning Meets Zero-Shot Learning: A Hybrid Approach for Fine-Grained Classification

Li Niu, Ashok Veeraraghavan, Ashutosh Sabharwal; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7171-7180

Fine-grained image classification, which targets at distinguishing subtle distinctions among various subordinate categories, remains a very difficult task due to the high annotation cost of enormous fine-grained categories. To cope with the scarcity of well-labeled training images, existing works mainly follow two research directions: 1) utilize freely available web images without human annotation; 2) only annotate some fine-grained categories and transfer the knowledge to other fine-grained categories, which falls into the scope of zero-shot learning (ZSL). However, the above two directions have their own drawbacks. For the first direction, the labels of web images are very noisy and the data distribution between web images and test images are considerably different. For the second direction, the performance gap between ZSL and traditional supervised learning is still very large. The drawbacks of the above two directions motivate us to design a new framework which can jointly leverage both web data and auxiliary labeled categories to predict the test categories that are not associated with any well-labeled training images. Comprehensive experiments on three benchmark datasets demonstrate the effectiveness of our proposed framework.

****

Look, Imagine and Match: Improving Textual-Visual Cross-Modal Retrieval With Generative Models

Jiuxiang Gu, Jianfei Cai, Shafiq R. Joty, Li Niu, Gang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7181-7189

Textual-visual cross-modal retrieval has been a hot research topic in both computer vision and natural language processing communities. Learning appropriate representations for multi-modal data is crucial for the cross-modal retrieval performance. Unlike existing image-text retrieval approaches that embed image-text pairs as single feature vectors in a common representational space, we propose to incorporate generative processes into the cross-modal feature embedding, through which we are able to learn not only the global abstract features but also the local grounded features. Extensive experiments show that our framework can well match images and sentences with complex content, and achieve the state-of-the-art cross-modal retrieval results on MSCOCO dataset.

****

Bidirectional Attentive Fusion With Context Gating for Dense Video Captioning

Jingwen Wang, Wenhao Jiang, Lin Ma, Wei Liu, Yong Xu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7190-7198

Dense video captioning is a newly emerging task that aims at both localizing and describing all events in a video. We identify and tackle two challenges on this task, namely, (1) how to utilize both past and future contexts for accurate eve

nt proposal predictions, and (2) how to construct informative input to the decoder for generating natural event descriptions. First, previous works predominantly generate temporal event proposals in the forward direction, which neglects future video contexts. We propose a bidirectional proposal method that effectively exploits both past and future contexts to make proposal predictions. Second, different events ending at (nearly) the same time are indistinguishable in the previous works, resulting in the same captions. We solve this problem by representing each event with an attentive fusion of hidden states from the proposal module and video contents (e.g., C3D features). We further propose a novel context gating mechanism to balance the contributions from the current event and its surrounding contexts dynamically. We empirically show that our attentively fused event representation is superior to the proposal hidden states or video contents alone. By coupling proposal and captioning modules into one unified framework, our model outperforms the state-of-the-arts on the ActivityNet Captions dataset with a relative gain of over 100% (Meteor score increases from 4.82 to 9.65).
****

InLoc: Indoor Visual Localization With Dense Matching and View Synthesis

Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, Akihiko Torii; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7199-7209

We seek to predict the 6 degree-of-freedom (6DoF) pose of a query photograph with respect to a large indoor 3D map. The contributions of this work are three-fold. First, we develop a new large-scale visual localization method targeted for indoor environments. The method proceeds along three steps: (i) efficient retrieval of candidate poses that ensures scalability to large-scale environments, (ii) pose estimation using dense matching rather than local features to deal with textureless indoor scenes, and (iii) pose verification by virtual view synthesis to cope with significant changes in viewpoint, scene layout, and occluders. Second, we collect a new dataset with reference 6DoF poses for large-scale indoor localization. Query photographs are captured by mobile phones at a different time than the reference 3D map, thus presenting a realistic indoor localization scenario. Third, we demonstrate that our method significantly outperforms current state-of-the-art indoor localization approaches on this new challenging data.
****

Towards High Performance Video Object Detection

Xizhou Zhu, Jifeng Dai, Lu Yuan, Yichen Wei; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7210-7218

There has been significant progresses for image object detection recently. Nevertheless, video object detection has received little attention, although it is more challenging and more important in practical scenarios. Built upon the recent works, this work proposes a unified viewpoint based on the principle of multi-frame end-to-end learning of features and cross-frame motion. Our approach extends prior works with three new techniques and steadily pushes forward the performance envelope (speed-accuracy tradeoff), towards high performance video object detection.
****

Neural Baby Talk

Jiasen Lu, Jianwei Yang, Dhruv Batra, Devi Parikh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7219-7228

We introduce a novel framework for image captioning that can produce natural language explicitly grounded in entities that object detectors find in the image. Our approach reconciles classical slot filling approaches (that are generally better grounded in images) with modern neural captioning approaches (that are generally more natural sounding and accurate). Our approach first generates a sentence `template' with slot locations explicitly tied to specific image regions. These slots are then filled in by visual concepts identified in the regions by object detectors. The entire architecture (sentence template generation and slot filling with object detectors) is end-to-end differentiable. We verify the effectiveness of our proposed model on different image captioning tasks. On standard image captioning and novel object captioning, our model reaches state-of-the-art on both COCO and Flickr30k datasets. We also demonstrate that our model has unique advantages when the train and test distributions of scene compositions -- and hence language priors of associated captions -- are different. Code has been made available at: https://github.com/jiasenlu/NeuralBabyTalk

****

Few-Shot Image Recognition by Predicting Parameters From Activations
Siyuan Qiao, Chenxi Liu, Wei Shen, Alan L. Yuille; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7229-7238
In this paper, we are interested in the few-shot learning problem. In particular, we focus on a challenging scenario where the number of categories is large and the number of examples per novel category is very limited, e.g. 1, 2, or 3. Motivated by the close relationship between the parameters and the activations in a neural network associated with the same category, we propose a novel method that can adapt a pre-trained neural network to novel categories by directly predicting the parameters from the activations. Zero training is required in adaptation to novel categories, and fast inference is realized by a single forward pass. We evaluate our method by doing few-shot image recognition on the ImageNet dataset, which achieves the state-of-the-art classification accuracy on novel categories by a significant margin while keeping comparable performance on the large-scale categories. We also test our method on the MiniImageNet dataset and it strongly outperforms the previous state-of-the-art methods.

****

Iterative Visual Reasoning Beyond Convolutions
Xinlei Chen, Li-Jia Li, Li Fei-Fei, Abhinav Gupta; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7239-7248
We present a novel framework for iterative visual reasoning. Our framework goes beyond current recognition systems that lack the capability to reason beyond stack of convolutions. The framework consists of two core modules: a local module that uses spatial memory to store previous beliefs in parallel; and a global graph-reasoning module. Our graph has three components: a) a knowledge graph where we represent classes as nodes and build edges to encode different types of semantic relationships between them; b) a region graph of the current image where regions in the image are nodes and spatial relationships between these regions are edges; c) an assignment graph that assigns regions to class nodes. Both the local module and the global module roll-out iteratively and cross-feed predictions to each other to refine estimates. The final predictions are made by combining the best of both modules with an attention mechanism. We show strong performance over plain ConvNets, eg achieving an $8.4%$ absolute improvement on ADE measured by per-class average precision. Analysis also shows that the framework is resilient to missing regions for reasoning.

****

Visual Question Reasoning on General Dependency Tree
Qingxing Cao, Xiaodan Liang, Bailing Li, Guanbin Li, Liang Lin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7249-7257
The collaborative reasoning for understanding each image-question pair is very critical but under-explored for an interpretable Visual Question Answering (VQA) system. Although very recent works also tried the explicit compositional processes to assemble multiple sub-tasks embedded in the questions, their models heavily rely on the annotations or hand-crafted rules to obtain valid reasoning layout, leading to either heavy labor or poor performance on composition reasoning. In this paper, to enable global context reasoning for better aligning image and language domains in diverse and unrestricted cases, we propose a novel reasoning network called Adversarial Composition Modular Network (ACMN). This network comprises of two collaborative modules: i) an adversarial attention module to exploit the local visual evidence for each word parsed from the question; ii) a residual composition module to compose the previously mined evidence. Given a dependency parse tree for each question, the adversarial attention module progressively discovers salient regions of one word by densely combining regions of child word nodes in an adversarial manner. Then residual composition module merges the hid

den representations of an arbitrary number of children through sum pooling and residual connection. Our ACMN is thus capable of building an interpretable VQA system that gradually dives the image cues following a question-driven reasoning route and makes global reasoning by incorporating the learned knowledge of all attention modules in a principled manner. Experiments on relational datasets demonstrate the superiority of our ACMN and visualization results show the explainable capability of our reasoning system.
****
CVM-Net: Cross-View Matching Network for Image-Based Ground-to-Aerial Geo-Localization
Sixing Hu, Mengdan Feng, Rang M. H. Nguyen, Gim Hee Lee; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7258-7267
The problem of localization on a geo-referenced aerial/satellite map given a query ground view image remains challenging due to the drastic change in viewpoint that causes traditional image descriptors based matching to fail. We leverage on the recent success of deep learning to propose the CVM-Net for the cross-view image-based ground-to-aerial geo-localization task. Specifically, our network is based on the Siamese architecture to do metric learning for the matching task. We first use the fully convolutional layers to extract local image features, which are then encoded into global image descriptors using the powerful NetVLAD. As part of the training procedure, we also introduce a simple yet effective weighted soft margin ranking loss function that not only speeds up the training convergence but also improves the final matching accuracy. Experimental results show that our proposed network significantly outperforms the state-of-the-art approaches on two existing benchmarking datasets.
****
Revisiting Dilated Convolution: A Simple Approach for Weakly- and Semi-Supervised Semantic Segmentation
Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, Thomas S. Huang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7268-7277
Despite remarkable progress, weakly supervised segmentation methods are still inferior to their fully supervised counterparts. We obverse that the performance gap mainly comes from the inability of producing dense and integral pixel-level object localization for training images only with image-level labels. In this work, we revisit the dilated convolution proposed in [1] and shed light on how it enables the classification network to generate dense object localization. By substantially enlarging the receptive fields of convolutional kernels with different dilation rates, the classification network can localize the object regions even when they are not so discriminative for classification and finally produce reliable object regions for benefiting both weakly- and semi- supervised semantic segmentation. Despite the apparent simplicity of dilated convolution, we are able to obtain superior performance for semantic segmentation tasks. In particular, it achieves 60.8% and 67.6% mean Intersection-over-Union (mIoU) on Pascal VOC 2012 test set in weakly- (only image-level labels are available) and semi- (1,464 segmentation masks are available) settings, which are the new state-of-the-arts.
****
Low-Shot Learning From Imaginary Data
Yu-Xiong Wang, Ross Girshick, Martial Hebert, Bharath Hariharan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7278-7286
Humans can quickly learn new visual concepts, perhaps because they can easily visualize or imagine what novel objects look like from different views. Incorporating this ability to hallucinate novel instances of new concepts might help machine vision systems perform better low-shot learning, i.e., learning concepts from few examples. We present a novel approach to low-shot learning that uses this idea. Our approach builds on recent progress in meta-learning (''learning to learn'') by combining a meta-learner with a ''hallucinator'' that produces additional training examples, and optimizing both models jointly. Our hallucinator can be

incorporated into a variety of meta-learners and provides significant gains: up to a 6 point boost in classification accuracy when only a single training example is available, yielding state-of-the-art performance on the challenging ImageNet low-shot classification benchmark.

****

DoubleFusion: Real-Time Capture of Human Performances With Inner Body Shapes From a Single Depth Sensor

Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, Yebin Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7287-7296

We propose DoubleFusion, a new real-time system that combines volumetric dynamic reconstruction with data-driven template fitting to simultaneously reconstruct detailed geometry, non-rigid motion and the inner human body shape from a single depth camera. One of the key contributions of this method is a double layer representation consisting of a complete parametric body shape inside and a gradually fused outer surface layer. A pre-defined node graph on the body surface parameterizes the non-rigid deformations near the body and a free-form dynamically changing graph parameterizes the outer surface layer far from the body allowing more general reconstruction. We further propose a joint motion tracking method based on the double layer representation to enable robust and fast motion tracking performance. Moreover, the inner body shape is optimized online and forced to fit inside the outer surface layer. Overall, our method enables increasingly denoised, detailed and complete surface reconstructions, fast motion tracking performance and plausible inner body shape reconstruction in real-time. In particular, experiments show improved fast motion tracking and loop closure performance on more challenging scenarios.

****

DensePose: Dense Human Pose Estimation in the Wild

R■za Alp Güler, Natalia Neverova, Iasonas Kokkinos; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7297-7306

In this work we establish dense correspondences between an RGB image and a surface-based representation of the human body, a task we refer to as dense human pose estimation. We gather dense correspondences for 50K persons appearing in the COCO dataset by introducing an efficient annotation pipeline. We then use our dataset to train CNN-based systems that deliver dense correspondence "in the wild", namely in the presence of background, occlusions and scale variations. We improve our training set's effectiveness by training an inpainting network that can fill in missing ground truth values and report improvements with respect to the best results that would be achievable in the past. We experiment with fully-convolutional networks and region-based models and observe a superiority of the latter. We further improve accuracy through cascading, obtaining a system that delivers highly-accurate results at multiple frames per second on a single gpu. Supplementary materials, data, code, and videos are provided on the project page http://densepose.org.

****

Ordinal Depth Supervision for 3D Human Pose Estimation

Georgios Pavlakos, Xiaowei Zhou, Kostas Daniilidis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7307-7316

Our ability to train end-to-end systems for 3D human pose estimation from single images is currently constrained by the limited availability of 3D annotations for natural images. Most datasets are captured using Motion Capture (MoCap) systems in a studio setting and it is difficult to reach the variability of 2D human pose datasets, like MPII or LSP. To alleviate the need for accurate 3D ground truth, we propose to use a weaker supervision signal provided by the ordinal depths of human joints. This information can be acquired by human annotators for a wide range of images and poses. We showcase the effectiveness and flexibility of training Convolutional Networks (ConvNets) with these ordinal relations in different settings, always achieving competitive performance with ConvNets trained with accurate 3D joint coordinates. Additionally, to demonstrate the potential of the approach, we augment the popular LSP and MPII datasets with ordinal depth ann otations. This extension allows us to present quantitative and qualitative evaluation in non-studio conditions. Simultaneously, these ordinal annotations can be easily incorporated in the training procedure of typical ConvNets for 3D human pose. Through this inclusion we achieve new state-of-the-art performance for the relevant benchmarks and validate the effectiveness of ordinal depth supervision for 3D human pose.

****

Consensus Maximization for Semantic Region Correspondences
Pablo Speciale, Danda P. Paudel, Martin R. Oswald, Hayko Riemenschneider, Luc Van Gool, Marc Pollefeys; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7317-7326

We propose a novel method for the geometric registration of semantically labeled regions. We approximate semantic regions by ellipsoids, and leverage their convexity to formulate the correspondence search effectively as a constrained optimization problem that maximizes the number of matched regions, and which we solve globally optimal in a branch-and-bound fashion. To this end, we derive suitable linear matrix inequality constraints which describe ellipsoid-to-ellipsoid assignment conditions. Our approach is robust to large percentages of outliers and thus applicable to difficult correspondence search problems. In multiple experiments we demonstrate the flexibility and robustness of our approach on a number of challenging vision problems.

****

Robust Hough Transform Based 3D Reconstruction From Circular Light Fields
Alessandro Vianello, Jens Ackermann, Maximilian Diebold, Bernd Jähne; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7327-7335

Light-field imaging is based on images taken on a regular grid. Thus, high-quality 3D reconstructions are obtainable by analyzing orientations in epipolar plane images (EPIs). Unfortunately, such data only allows to evaluate one side of the object. Moreover, a constant intensity along each orientation is mandatory for most of the approaches. This paper presents a novel method which allows to reconstruct depth information from data acquired with a circular camera motion, termed circular light fields. With this approach it is possible to determine the full 360 degree view of target objects. Additionally, circular light fields allow retrieving depth from datasets acquired with telecentric lenses, which is not possible with linear light fields. The proposed method finds trajectories of 3D points in the EPIs by means of a modified Hough transform. For this purpose, binary EPI-edge images are used, which not only allow to obtain reliable depth information, but also overcome the limitation of constant intensity along trajectories. Experimental results on synthetic and real datasets demonstrate the quality of the proposed algorithm.

****

Alive Caricature From 2D to 3D
Qianyi Wu, Juyong Zhang, Yu-Kun Lai, Jianmin Zheng, Jianfei Cai; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7336-7345

Caricature is an art form that expresses subjects in abstract, simple and exaggerated views. While many caricatures are 2D images, this paper presents an algorithm for creating expressive 3D caricatures from 2D caricature images with minimum user interaction. The key idea of our approach is to introduce an intrinsic deformation representation that has the capability of extrapolation, enabling us to create a deformation space from standard face datasets, which maintains face constraints and meanwhile is sufficiently large for producing exaggerated face models. Built upon the proposed deformation representation, an optimization model is formulated to find the 3D caricature that captures the style of the 2D caricature image automatically. The experiments show that our approach has better capability in expressing caricatures than those fitting approaches directly using classical parametric face models such as 3DMM and FaceWareHouse. Moreover, our approach is based on standard face datasets and avoids constructing complicated 3D caricature training sets, which provides great flexibility in real applications.

****

Nonlinear 3D Face Morphable Model

Luan Tran, Xiaoming Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7346-7355

As a classic statistical model of 3D facial shape and texture, 3D Morphable Model (3DMM) is widely used in facial analysis, e.g., model fitting, image synthesis. Conventional 3DMM is learned from a set of well-controlled 2D face images with associated 3D face scans, and represented by two sets of PCA basis functions. Due to the type and amount of training data, as well as the linear bases, the representation power of 3DMM can be limited. To address these problems, this paper proposes an innovative framework to learn a nonlinear 3DMM model from a large set of unconstrained face images, without collecting 3D face scans. Specifically, given a face image as input, a network encoder estimates the projection, shape and texture parameters. Two decoders serve as the nonlinear 3DMM to map from the shape and texture parameters to the 3D shape and texture, respectively. With the projection parameter, 3D shape, and texture, a novel analytically-differentiable rendering layer is designed to reconstruct the original input face. The entire network is end-to-end trainable with only weak supervision. We demonstrate the superior representation power of our nonlinear 3DMM over its linear counterpart, and its contribution to face alignment and 3D reconstruction.

****

Through-Wall Human Pose Estimation Using Radio Signals

Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, Dina Katabi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7356-7365

This paper demonstrates accurate human pose estimation through walls and occlusions. We leverage the fact that wireless signals in the WiFi frequencies traverse walls and reflect off the human body. We introduce a deep neural network approach that parses such radio signals to estimate 2D poses. Since humans cannot annotate radio signals, we use state-of-the-art vision model to provide cross-modal supervision. Specifically, during training the system uses synchronized wireless and visual inputs, extracts pose information from the visual stream, and uses it to guide the training process. Once trained, the network uses only the wireless signal for pose estimation. We show that, when tested on visible scenes, the radio-based system is almost as accurate as the vision-based system used to train it. Yet, unlike vision-based pose estimation, the radio-based system can estimate 2D poses through walls despite never trained on such scenarios. Demo videos are available at our website (http://rfpose.csail.mit.edu).

****

What Makes a Video a Video: Analyzing Temporal Information in Video Understanding Models and Datasets

De-An Huang, Vignesh Ramanathan, Dhruv Mahajan, Lorenzo Torresani, Manohar Paluri, Li Fei-Fei, Juan Carlos Niebles; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7366-7375

The ability to capture temporal information has been critical to the development of video understanding models. While there have been numerous attempts at modeling motion in videos, an explicit analysis of the effect of temporal information for video understanding is still missing. In this work, we aim to bridge this gap and ask the following question: How important is the motion in the video for recognizing the action? To this end, we propose two novel frameworks: (i) class-agnostic temporal generator and (ii) motion-invariant frame selector to reduce/remove motion for an ablation analysis without introducing other artifacts. This isolates the analysis of motion from other aspects of the video. The proposed frameworks provide a much tighter estimate of the effect of motion (from 25% to 6% on UCF101 and 15% to 5% on Kinetics) compared to baselines in our analysis. Our analysis provides critical insights about existing models like C3D, and how it could be made to achieve comparable results with a sparser set of frames.

****

Fast Video Object Segmentation by Reference-Guided Mask Propagation

Seoung Wug Oh, Joon-Young Lee, Kalyan Sunkavalli, Seon Joo Kim; Proceedings of t he IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7376-7385

We present an efficient method for the semi-supervised video object segmentation. Our method achieves accuracy competitive with state-of-the-art methods while running in a fraction of time compared to others. To this end, we propose a deep Siamese encoder-decoder network that is designed to take advantage of mask propagation and object detection while avoiding the weaknesses of both approaches. Our network, learned through a two-stage training process that exploits both synthetic and real data, works robustly without any online learning or post-processing. We validate our method on four benchmark sets that cover single and multiple object segmentation. On all the benchmark sets, our method shows comparable accuracy while having the order of magnitude faster runtime. We also provide extensive ablation and add-on studies to analyze and evaluate our framework.

****

NeuralNetwork-Viterbi: A Framework for Weakly Supervised Video Learning

Alexander Richard, Hilde Kuehne, Ahsan Iqbal, Juergen Gall; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7386-7395

Video learning is an important task in computer vision and has experienced increasing interest over the recent years. Since even a small amount of videos easily comprises several million frames, methods that do not rely on a frame-level annotation are of special importance. In this work, we propose a novel learning algorithm with a Viterbi-based loss that allows for online and incremental learning of weakly annotated video data. We moreover show that explicit context and length modeling leads to huge improvements in video segmentation and labeling tasks and include these models into our framework. On several action segmentation benchmarks, we obtain an improvement of up to 10% compared to current state-of-the-art methods.

****

Actor and Observer: Joint Modeling of First and Third-Person Videos

Gunnar A. Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, Karteek Alahari; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7396-7404

Several theories in cognitive neuroscience suggest that when people interact with the world, or simulate interactions, they do so from a first-person egocentric perspective, and seamlessly transfer knowledge between third-person (observer) and first-person (actor). Despite this, learning such models for human action recognition has not been achievable due to the lack of data. This paper takes a step in this direction, with the introduction of Charades-Ego, a large-scale dataset of paired first-person and third-person videos, involving 112 people, with 4000 paired videos. This enables learning the link between the two, actor and observer perspectives. Thereby, we address one of the biggest bottlenecks facing egocentric vision research, providing a link from first-person to the abundant third-person data on the web. We use this data to learn a joint representation of first and third-person videos, with only weak supervision, and show its effectiveness for transferring knowledge from the third-person to the first-person domain.

****

HSA-RNN: Hierarchical Structure-Adaptive RNN for Video Summarization

Bin Zhao, Xuelong Li, Xiaoqiang Lu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7405-7414

Although video summarization has achieved great success in recent years, few approaches have realized the influence of video structure on the summarization results. As we know, the video data follow a hierarchical structure, i.e., a video is composed of shots, and a shot is composed of several frames. Generally, shots provide the activity-level information for people to understand the video content. While few existing summarization approaches pay attention to the shot segmentation procedure. They generate shots by some trivial strategies, such as fixed length segmentation, which may destroy the underlying hierarchical structure of video data and further reduce the quality of generated summaries. To address this problem, we propose a structure-adaptive video summarization approach that inte grates shot segmentation and video summarization into a Hierarchical Structure-Adaptive RNN, denoted as HSA-RNN. We evaluate the proposed approach on four popular datasets, i.e., SumMe, TVsum, CoSum and VTW. The experimental results have demonstrated the effectiveness of HSA-RNN in the video summarization task.

****

Fast and Accurate Online Video Object Segmentation via Tracking Parts

Jingchun Cheng, Yi-Hsuan Tsai, Wei-Chih Hung, Shengjin Wang, Ming-Hsuan Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7415-7424

Online video object segmentation is a challenging task as it entails to process the image sequence timely and accurately. To segment a target object through the video, numerous CNN-based methods have been developed by heavily finetuning on the object mask in the first frame, which is time-consuming for online applications. In this paper, we propose a fast and accurate video object segmentation algorithm that can immediately start the segmentation process once receiving the images. We first utilize a part-based tracking method to deal with challenging factors such as large deformation, occlusion, and cluttered background. Based on the tracked bounding boxes of parts, we construct a region-of-interest segmentation network to generate part masks. Finally, a similarity-based scoring function is adopted to refine these object parts by comparing them to the visual information in the first frame. Our method performs favorably against state-of-the-art algorithms in accuracy on the DAVIS benchmark dataset, while achieving much faster runtime performance.

****

Now You Shake Me: Towards Automatic 4D Cinema

Yuhao Zhou, Makarand Tapaswi, Sanja Fidler; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7425-7434

We are interested in enabling automatic 4D cinema by parsing physical and special effects from untrimmed movies. These include effects such as physical interactions, water splashing, light, and shaking, and are grounded to either a character in the scene or the camera. We collect a new dataset referred to as the Movie4D dataset which annotates over 9K effects in 63 movies. We propose a Conditional Random Field model atop a neural network that brings together visual and audio information, as well as semantics in the form of person tracks. Our model further exploits correlations of effects between different characters in the clip as well as across movie threads. We propose effect detection and classification as two tasks, and present results along with ablation studies on our dataset, paving the way towards 4D cinema in everyone’s homes.

****

Viewpoint-Aware Video Summarization

Atsushi Kanehira, Luc Van Gool, Yoshitaka Ushiku, Tatsuya Harada; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7435-7444

This paper introduces a novel variant of video summarization, namely building a summary that depends on the particular aspect of a video the viewer focuses on. We refer to this as viewpoint. To infer what the desired viewpoint may be, we assume that several other videos are available, especially groups of videos, e.g., as folders on a person's phone or laptop. The semantic similarity between videos in a group vs. the dissimilarity between groups is used to produce viewpoint-specific summaries. For considering similarity as well as avoiding redundancy, output summary should be (A) diverse, (B) representative of videos in the same group, and (C) discriminative against videos in the different groups. To satisfy these requirements (A)-(C) simultaneously, we proposed a novel video summarization method from multiple groups of videos. Inspired by Fisher's discriminant criteria, it selects summary by optimizing the combination of three terms (a) inner-summary, (b) inner-group, and (c) between-group variances defined on the feature representation of summary, which can simply represent (A)-(C). Moreover, we developed a novel dataset to investigate how well the generated summary reflects the underlying viewpoint. Quantitative and qualitative experiments conducted on the dataset demonstrate the effectiveness of proposed method.

****

Photometric Stereo in Participating Media Considering Shape-Dependent Forward Scatter

Yuki Fujimura, Masaaki Iiyama, Atsushi Hashimoto, Michihiko Minoh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7445-7453

Images captured in participating media such as murky water, fog, or smoke are degraded by scattered light. Thus, the use of traditional three-dimensional (3D) reconstruction techniques in such environments is difficult. In this paper, we propose a photometric stereo method for participating media. The proposed method differs from previous studies with respect to modeling shape-dependent forward scatter. In the proposed model, forward scatter is described as an analytical form using lookup tables and is represented by spatially-variant kernels. We also propose an approximation of a large-scale dense matrix as a sparse matrix, which enables the removal of forward scatter. Experiments with real and synthesized data demonstrate that the proposed method improves 3D reconstruction in participating media.

****

Direction-Aware Spatial Context Features for Shadow Detection

Xiaowei Hu, Lei Zhu, Chi-Wing Fu, Jing Qin, Pheng-Ann Heng; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7454-7462

Shadow detection is a fundamental and challenging task, since it requires an understanding of global image semantics and there are various backgrounds around shadows. This paper presents a novel network for shadow detection by analyzing image context in a direction-aware manner. To achieve this, we first formulate the direction-aware attention mechanism in a spatial recurrent neural network (RNN) by introducing attention weights when aggregating spatial context features in the RNN. By learning these weights through training, we can recover direction-aware spatial context (DSC) for detecting shadows. This design is developed into the DSC module and embedded in a CNN to learn DSC features at different levels. Moreover, a weighted cross entropy loss is designed to make the training more effective. We employ two common shadow detection benchmark datasets and perform various experiments to evaluate our network. Experimental results show that our network outperforms state-of-the-art methods and achieves 97% accuracy and 38% reduction on balance error rate.

****

Discriminative Learning of Latent Features for Zero-Shot Recognition

Yan Li, Junge Zhang, Jianguo Zhang, Kaiqi Huang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7463-7471

Zero-shot learning (ZSL) aims to recognize unseen image categories by learning an embedding space between image and semantic representations. For years, among existing works, it has been the center task to learn the proper mapping matrices aligning the visual and semantic space, whilst the importance to learn discriminative representations for ZSL is ignored. In this work, we retrospect existing methods and demonstrate the necessity to learn discriminative representations for both visual and semantic instances of ZSL. We propose an end-to-end network that is capable of 1) automatically discovering discriminative regions by a zoom network; and 2) learning discriminative semantic representations in an augmented space introduced for both user-defined and latent attributes. Our proposed method is tested extensively on two challenging ZSL datasets, and the experiment results show that the proposed method significantly outperforms state-of-the-art methods.

****

Learning to Adapt Structured Output Space for Semantic Segmentation

Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, Manmohan Chandraker; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7472-7481

Convolutional neural network-based approaches for semantic segmentation rely on supervision with pixel-level ground truth, but may not generalize well to unseen image domains. As the labeling process is tedious and labor intensive, developing algorithms that can adapt source ground truth labels to the target domain is of great interest. In this paper, we propose an adversarial learning method for domain adaptation in the context of semantic segmentation. Considering semantic segmentations as structured outputs that contain spatial similarities between the source and target domains, we adopt adversarial learning in the output space. To further enhance the adapted model, we construct a multi-level adversarial network to effectively perform output space domain adaptation at different feature levels. To further improve our method, we utilize multi-level output adaptation based on feature maps at different levels. Extensive experiments and ablation study are conducted under various domain adaptation settings, including synthetic-to-real and cross-city scenarios. We show that the proposed method performs favorably against the state-of-the-art methods in terms of accuracy and visual quality.

****

Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics

Alex Kendall, Yarin Gal, Roberto Cipolla; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7482-7491

Numerous deep learning applications benefit from multi-task learning with multiple regression and classification objectives. In this paper we make the observation that the performance of such systems is strongly dependent on the relative weighting between each task's loss. Tuning these weights by hand is a difficult and expensive process, making multi-task learning prohibitive in practice. We propose a principled approach to multi-task deep learning which weighs multiple loss functions by considering the homoscedastic uncertainty of each task. This allows us to simultaneously learn various quantities with different units or scales in both classification and regression settings. We demonstrate our model learning per-pixel depth regression, semantic and instance segmentation from a monocular input image. Perhaps surprisingly, we show our model can learn multi-task weightings and outperform separate models trained individually on each task.

****

Jointly Localizing and Describing Events for Dense Video Captioning

Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, Tao Mei; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7492-7500

Automatically describing a video with natural language is regarded as a fundamental challenge in computer vision. The problem nevertheless is not trivial especially when a video contains multiple events to be worthy of mention, which often happens in real videos. A valid question is how to temporally localize and then describe events, which is known as ``dense video captioning." In this paper, we present a novel framework for dense video captioning that unifies the localization of temporal event proposals and sentence generation of each proposal, by jointly training them in an end-to-end manner. To combine these two worlds, we integrate a new design, namely descriptiveness regression, into a single shot detection structure to infer the descriptive complexity of each detected proposal via sentence generation. This in turn adjusts the temporal locations of each event proposal. Our model differs from existing dense video captioning methods since we propose a joint and global optimization of detection and captioning, and the framework uniquely capitalizes on an attribute-augmented video captioning architecture. Extensive experiments are conducted on ActivityNet Captions dataset and our framework shows clear improvements when compared to the state-of-the-art techniques. More remarkably, we obtain a new record: METEOR of 12.96% on ActivityNet Captions official test set.

****

Going From Image to Video Saliency: Augmenting Image Salience With Dynamic Attentional Push

Siavash Gorji, James J. Clark; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7501-7511

We present a novel method to incorporate the recent advent in static saliency mo

dels to predict the saliency in videos. Our model augments the static saliency models with the Attentional Push effect of the photographer and the scene actors in a shared attention setting. We demonstrate that not only it is imperative to use static Attentional Push cues, noticeable performance improvement is achievable by learning the time-varying nature of Attentional Push. We propose a multi-stream Convolutional Long Short-Term Memory network (ConvLSTM) structure which augments state-of-the-art in static saliency models with dynamic Attentional Push. Our network contains four pathways, a saliency pathway and three Attentional Push pathways. The multi-pathway structure is followed by an augmenting convnet that learns to combine the complementary and time-varying outputs of the ConvLSTMs by minimizing the relative entropy between the augmented saliency and viewers fixation patterns on videos. We evaluate our model by comparing the performance of several augmented static saliency models with state-of-the-art in spatiotemporal saliency on three largest dynamic eye tracking datasets, HOLLYWOOD2, UCF-Sport and DIEM. Experimental results illustrates that solid performance gain is achievable using the proposed methodology.
****

M3: Multimodal Memory Modelling for Video Captioning

Junbo Wang, Wei Wang, Yan Huang, Liang Wang, Tieniu Tan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7512-7520

Video captioning which automatically translates video clips into natural language sentences is a very important task in computer vision. By virtue of recent deep learning technologies, video captioning has made great progress. However, learning an effective mapping from the visual sequence space to the language space is still a challenging problem due to the long-term multimodal dependency modelling and semantic misalignment. Inspired by the facts that memory modelling poses potential advantages to long-term sequential problems [35] and working memory is the key factor of visual attention [33], we propose a Multimodal Memory Model (M3) to describe videos, which builds a visual and textual shared memory to model the long-term visual-textual dependency and further guide visual attention on described visual targets to solve visual-textual alignments. Specifically, similar to [10], the proposed M3 attaches an external memory to store and retrieve both visual and textual contents by interacting with video and sentence with multiple read and write operations. To evaluate the proposed model, we perform experiments on two public datasets: MSVD and MSR-VTT. The experimental results demonstrate that our method outperforms most of the state-of-the-art methods in terms of BLEU and METEOR.
****

Emotional Attention: A Study of Image Sentiment and Visual Attention

Shaojing Fan, Zhiqi Shen, Ming Jiang, Bryan L. Koenig, Juan Xu, Mohan S. Kankanhalli, Qi Zhao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7521-7531

Image sentiment influences visual perception. Emotion-eliciting stimuli such as happy faces and poisonous snakes are generally prioritized in human attention. However, little research has evaluated the interrelationships of image sentiment and visual saliency. In this paper, we present the first study to focus on the relation between emotional properties of an image and visual attention. We first create the EMOtional attention dataset (EMOd). It is a diverse set of emotion-eliciting images, and each image has (1) eye-tracking data collected from 16 subjects, (2) intensive image context labels including object contour, object sentiment, object semantic category, and high-level perceptual attributes such as image aesthetics and elicited emotions. We perform extensive analyses on EMOd to identify how image sentiment relates to human attention. We discover an emotion prioritization effect: for our images, emotion-eliciting content attracts human attention strongly, but such advantage diminishes dramatically after initial fixation. Aiming to model the human emotion prioritization computationally, we design a deep neural network for saliency prediction, which includes a novel subnetwork that learns the spatial and semantic context of the image scene. The proposed network outperforms the state-of-the-art on three benchmark datasets, by effective ly capturing the relative importance of human attention within an image. The code, models, and dataset are available online at https://nus-sesame.top/emotionalattention/.

****

A Low Power, High Throughput, Fully Event-Based Stereo System

Alexander Andreopoulos, Hirak J. Kashyap, Tapan K. Nayak, Arnon Amir, Myron D. Flickner; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7532-7542

We introduce a stereo correspondence system implemented fully on event-based digital hardware, using a fully graph-based non von-Neumann computation model, where no frames, arrays, or any other such data-structures are used. This is the first time that an end-to-end stereo pipeline from image acquisition and rectification, multi-scale spatio-temporal stereo correspondence, winner-take-all, to disparity regularization is implemented fully on event-based hardware. Using a cluster of TrueNorth neurosynaptic processors, we demonstrate their ability to process bilateral event-based inputs streamed live by Dynamic Vision Sensors (DVS), at up to 2,000 disparity maps per second, producing high fidelity disparities which are in turn used to reconstruct, at low power, the depth of events produced from rapidly changing scenes. Experiments on real-world sequences demonstrate the ability of the system to take full advantage of the asynchronous and sparse nature of DVS sensors for low power depth reconstruction, in environments where conventional frame-based cameras connected to synchronous processors would be inefficient for rapidly moving objects. System evaluation on event-based sequences demonstrates a ~200X improvement in terms of power per pixel per disparity map compared to the closest state-of-the-art, and maximum latencies of up to 11ms from spike injection to disparity map ejection.

****

VITON: An Image-Based Virtual Try-On Network

Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, Larry S. Davis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7543-7552

We present an image-based VIirtual Try-On Network (VITON) without using 3D information in any form, which seamlessly transfers a desired clothing item onto the corresponding region of a person using a coarse-to-fine strategy. Conditioned upon a new clothing-agnostic yet descriptive person representation, our framework first generates a coarse synthesized image with the target clothing item overlaid on that same person in the same pose. We further enhance the initial blurry clothing area with a refinement network. The network is trained to learn how much detail to utilize from the target clothing item, and where to apply to the person in order to synthesize a photo-realistic image in which the target item deforms naturally with clear visual patterns. Experiments on our newly collected Zalando dataset demonstrate its promise in the image-based virtual try-on task over state-of-the-art generative models.

****

Multi-Oriented Scene Text Detection via Corner Localization and Region Segmentation

Pengyuan Lyu, Cong Yao, Wenhao Wu, Shuicheng Yan, Xiang Bai; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7553-7563

Previous deep learning based state-of-the-art scene text detection methods can be roughly classified into two categories. The first category treats scene text as a type of general objects and follows general object detection paradigm to localize scene text by regressing the text box locations, but troubled by the arbitrary-orientation and large aspect ratios of scene text. The second one segments text regions directly, but mostly needs complex post processing. In this paper, we present a method that combines the ideas of the two types of methods while avoiding their shortcomings. We propose to detect scene text by localizing corner points of text bounding boxes and segmenting text regions in relative positions. In inference stage, candidate boxes are generated by sampling and grouping corner points, which are further scored by segmentation maps and suppressed by NMS.

Compared with previous methods, our method can handle long oriented text naturally and doesn’t need complex post processing. The experiments on ICDAR2013, ICDAR2015, MSRA-TD500, MLT and COCO-Text demonstrate that the proposed algorithm achieves better or comparable results in both accuracy and efficiency. Based on VGG16, it achieves an F-measure of 84:3% on ICDAR2015 and 81:5% on MSRA-TD500.

****

Multi-Content GAN for Few-Shot Font Style Transfer

Samaneh Azadi, Matthew Fisher, Vladimir G. Kim, Zhaowen Wang, Eli Shechtman, Trevor Darrell; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7564-7573

In this work, we focus on the challenge of taking partial observations of highly-stylized text and generalizing the observations to generate unobserved glyphs in the ornamented typeface. To generate a set of multi-content images following a consistent style from very few examples, we propose an end-to-end stacked conditional GAN model considering content along channels and style along network layers. Our proposed network transfers the style of given glyphs to the contents of unseen ones, capturing highly stylized fonts found in the real-world such as those on movie posters or infographics. We seek to transfer both the typographic stylization (ex. serifs and ears) as well as the textual stylization (ex. color gradients and effects.) We base our experiments on our collected data set including 10,000 fonts with different styles and demonstrate effective generalization from a very small number of observed glyphs.

****

Audio to Body Dynamics

Eli Shlizerman, Lucio Dery, Hayden Schoen, Ira Kemelmacher-Shlizerman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7574-7583

We present a method that gets as input an audio of violin or piano playing, and outputs a video of skeleton predictions which are further used to animate an avatar. The key idea is to create an animation of an avatar that moves their hands similarly to how a pianist or violinist would do, just from audio. Notably, it's not clear if body movement can be predicted from music at all and our aim in this work is to explore this possibility. In this paper, we present the first result that shows that natural body dynamics can be predicted. We built an LSTM network that is trained on violin and piano recital videos uploaded to the Internet. The predicted points are applied onto a rigged avatar to create the animation.

****

Weakly Supervised Coupled Networks for Visual Sentiment Analysis

Jufeng Yang, Dongyu She, Yu-Kun Lai, Paul L. Rosin, Ming-Hsuan Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7584-7592

Automatic assessment of sentiment from visual content has gained considerable attention with the increasing tendency of expressing opinions on-line. In this paper, we solve the problem of visual sentiment analysis using the high-level abstraction in the recognition process. Existing methods based on convolutional neural networks learn sentiment representations from the holistic image appearance. However, different image regions can have a different influence on the intended expression. This paper presents a weakly supervised coupled convolutional network with two branches to leverage the localized information. The first branch detects a sentiment specific soft map by training a fully convolutional network with the cross spatial pooling strategy, which only requires image-level labels, thereby significantly reducing the annotation burden. The second branch utilizes both the holistic and localized information by coupling the sentiment map with deep features for robust classification. We integrate the sentiment detection and classification branches into a unified deep framework and optimize the network in an end-to-end manner. Extensive experiments on six benchmark datasets demonstrate that the proposed method performs favorably against the state-ofthe-art methods for visual sentiment analysis.

****

Future Person Localization in First-Person Videos

Takuma Yagi, Karttikeya Mangalam, Ryo Yonetani, Yoichi Sato; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7593-7602

We present a new task that predicts future locations of people observed in first-person videos. Consider a first-person video stream continuously recorded by a wearable camera. Given a short clip of a person that is extracted from the complete stream, we aim to predict that person's location in future frames. To facilitate this future person localization ability, we make the following three key observations: a) First-person videos typically involve significant ego-motion which greatly affects the location of the target person in future frames; b) Scales of the target person act as a salient cue to estimate a perspective effect in first-person videos; c) First-person videos often capture people up-close, making it easier to leverage target poses (e.g., where they look) for predicting their future locations. We incorporate these three observations into a prediction framework with a multi-stream convolution-deconvolution architecture. Experimental results reveal our method to be effective on our new dataset as well as on a public social interaction dataset.

****

Preserving Semantic Relations for Zero-Shot Learning

Yashas Annadani, Soma Biswas; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7603-7612

Zero-shot learning has gained popularity due to its potential to scale recognition models without requiring additional training data. This is usually achieved by associating categories with their semantic information like attributes. However, we believe that the potential offered by this paradigm is not yet fully exploited. In this work, we propose to utilize the structure of the space spanned by the attributes using a set of relations. We devise objective functions to preserve these relations in the embedding space, thereby inducing semanticity to the embedding space. Through extensive experimental evaluation on five benchmark datasets, we demonstrate that inducing semanticity to the embedding space is beneficial for zero-shot learning. The proposed approach outperforms the state-of-the-art on the standard zero-shot setting as well as the more realistic generalized zero-shot setting. We also demonstrate how the proposed approach can be useful for making approximate semantic inferences about an image belonging to a category for which attribute information is not available.

****

Show Me a Story: Towards Coherent Neural Story Illustration

Hareesh Ravi, Lezi Wang, Carlos Muniz, Leonid Sigal, Dimitris Metaxas, Mubbasir Kapadia; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7613-7621

We propose an end-to-end network for the visual illustration of a sequence of sentences forming a story. At the core of our model is the ability to model the inter-related nature of the sentences within a story, as well as the ability to learn coherence to support reference resolution. The framework takes the form of an encoder-decoder architecture, where sentences are encoded using a hierarchical two-level sentence-story GRU, combined with an encoding of coherence, and sequentially decoded using predicted feature representation into a consistent illustrative image sequence. We optimize all parameters of our network in an end-to-end fashion with respect to order embedding loss, encoding entailment between images and sentences. Experiments on the VIST storytelling dataset cite{vist} highlight the importance of our algorithmic choices and efficacy of our overall model.

****

Reconstruction Network for Video Captioning

Bairui Wang, Lin Ma, Wei Zhang, Wei Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7622-7631

In this paper, the problem of describing visual contents of a video sequence with natural language is addressed. Unlike previous video captioning work mainly exploiting the cues of video contents to make a language description, we propose a reconstruction network (RecNet) with a novel encoder-decoder-reconstructor arch

itecture, which leverages both the forward (video to sentence) and backward (sentence to video) flows for video captioning. Specifically, the encoder-decoder makes use of the forward flow to produce the sentence description based on the encoded video semantic features. Two types of reconstructors are customized to employ the backward flow and reproduce the video features based on the hidden state sequence generated by the decoder. The generation loss yielded by encoder-decoder and the reconstruction loss introduced by reconstructor are jointly drawn into training the proposed RecNet in an end-to-end fashion. Experimental results on benchmark datasets demonstrate that the proposed reconstructor could boost the encoder-decoder models and leads to significant gains on video caption accuracy.
****

Fast Spectral Ranking for Similarity Search
Ahmet Iscen, Yannis Avrithis, Giorgos Tolias, Teddy Furon, Ond■ej Chum; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7632-7641

Despite the success of deep learning on representing images for particular object retrieval, recent studies show that the learned representations still lie on manifolds in a high dimensional space. This makes the Euclidean nearest neighbor search biased for this task. Exploring the manifolds online remains expensive even if a nearest neighbor graph has been computed offline. This work introduces an explicit embedding reducing manifold search to Euclidean search followed by dot product similarity search. This is equivalent to linear graph filtering of a sparse signal in the frequency domain. To speed up online search, we compute an approximate Fourier basis of the graph offline. We improve the state of art on particular object retrieval datasets including the challenging Instre dataset containing small objects. At a scale of 10^5 images, the offline cost is only a few hours, while query time is comparable to standard similarity search.
****

Mining on Manifolds: Metric Learning Without Labels
Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, Ond■ej Chum; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7642-7651

In this work we present a novel unsupervised framework for hard training example mining. The only input to the method is a collection of images relevant to the target application and a meaningful initial representation, provided e.g. by pre-trained CNN. Positive examples are distant points on a single manifold, while negative examples are nearby points on different manifolds. Both types of examples are revealed by disagreements between Euclidean and manifold similarities. The discovered examples can be used in training with any discriminative loss.   The method is applied to unsupervised fine-tuning of pre-trained networks for fine-grained classification and particular object retrieval. Our models are on par or are outperforming prior models that are fully or partially supervised.
****

PIXOR: Real-Time 3D Object Detection From Point Clouds
Bin Yang, Wenjie Luo, Raquel Urtasun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7652-7660

We address the problem of real-time 3D object detection from point clouds in the context of autonomous driving. Speed is critical as detection is a necessary component for safety. Existing approaches are, however, expensive in computation due to high dimensionality of point clouds. We utilize the 3D data more efficiently by representing the scene from the Bird's Eye View (BEV), and propose PIXOR, a proposal-free, single-stage detector that outputs oriented 3D object estimates decoded from pixel-wise neural network predictions. The input representation, network architecture, and model optimization are specially designed to balance high accuracy and real-time efficiency. We validate PIXOR on two datasets: the KITTI BEV object detection benchmark, and a large-scale 3D vehicle detection benchmark. In both datasets we show that the proposed detector surpasses other state-of-the-art methods notably in terms of Average Precision (AP), while still runs at 10 FPS.
****

Leveraging Unlabeled Data for Crowd Counting by Learning to Rank

Xialei Liu, Joost van de Weijer, Andrew D. Bagdanov; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7661-7669

We propose a novel crowd counting approach that leverages abundantly available unlabeled crowd imagery in a learning-to-rank framework. To induce a ranking of cropped images , we use the observation that any sub-image of a crowded scene image is guaranteed to contain the same number or fewer persons than the super-image. This allows us to address the problem of limited size of existing datasets for crowd counting. We collect two crowd scene datasets from Google using keyword searches and query-by-example image retrieval, respectively. We demonstrate how to efficiently learn from these unlabeled datasets by incorporating learning-to-rank in a multi-task network which simultaneously ranks images and estimates crowd density maps. Experiments on two of the most challenging crowd counting datasets show that our approach obtains state-of-the-art results.

****

Zero-Shot Kernel Learning

Hongguang Zhang, Piotr Koniusz; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7670-7679

In this paper, we address an open problem of zero-shot learning. Its principle is based on learning a mapping that associates feature vectors extracted from i.e. images and attribute vectors that describe objects and/or scenes of interest. In turns, this allows classifying unseen object classes and/or scenes by matching feature vectors via mapping to a newly defined attribute vector describing a new class. Due to importance of such a learning task, there exist many methods that learn semantic, probabilistic, linear or piece-wise linear mappings. In contrast, we apply well-established kernel methods to learn a non-linear mapping between the feature and attribute spaces. We propose an easy learning objective with orthogonality constraints inspired by the Linear Discriminant Analysis, Kernel-Target Alignment and Kernel Polarization methods. We evaluate the performance of our algorithm on the Polynomial as well as shift-invariant Gaussian and Cauchy kernels. Despite simplicity of our approach, we obtain state-of-the-art results on several zero-shot learning datasets and benchmarks including very recent AWA2 dataset.

****

Differential Attention for Visual Question Answering

Badri Patro, Vinay P. Namboodiri; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7680-7688

In this paper we aim to answer questions based on images when provided with a dataset of question-answer pairs for a number of images during training. A number of methods have focused on solving this problem by using image based attention. This is done by focusing on a specific part of the image while answering the question. Humans also do so when solving this problem. However, the regions that the previous systems focus on are not correlated with the regions that humans focus on. The accuracy is limited due to this drawback. In this paper, we propose to solve this problem by using an exemplar based method. We obtain one or more supporting and opposing exemplars to obtain a differential attention region. This differential attention is closer to human attention than other image based attention methods. It also helps in obtaining improved accuracy when answering questions. The method is evaluated on challenging benchmark datasets. We perform better than other image based attention methods and are competitive with other state of the art methods that focus on both image and questions.

****

Learning From Noisy Web Data With Category-Level Supervision

Li Niu, Qingtao Tang, Ashok Veeraraghavan, Ashutosh Sabharwal; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7689-7698

Learning from web data is increasingly popular due to abundant free web resources. However, the performance gap between webly supervised learning and traditional supervised learning is still very large, due to the label noise of web data. T o fill this gap, most existing methods propose to purify or augment web data using instance-level supervision, which generally requires heavy annotation. Instead, we propose to address the label noise by using more accessible category-level supervision. In particular, we build our deep probabilistic framework upon variational autoencoder (VAE), in which classification network and VAE can jointly leverage category-level hybrid information. Extensive experiments on three benchmark datasets demonstrate the effectiveness of our proposed method.

****

Toward Driving Scene Understanding: A Dataset for Learning Driver Behavior and Causal Reasoning

Vasili Ramanishka, Yi-Ting Chen, Teruhisa Misu, Kate Saenko; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7699-7707

Driving Scene understanding is a key ingredient for intelligent transportation systems. To achieve systems that can operate in a complex physical and social environment, they need to understand and learn how humans drive and interact with traffic scenes. We present the Honda Research Institute Driving Dataset (HDD), a challenging dataset to enable research on learning driver behavior in real-life environments. The dataset includes 104 hours of real human driving in the San Francisco Bay Area collected using an instrumented vehicle equipped with different sensors. We provide a detailed analysis of HDD with a comparison to other driving datasets. A novel annotation methodology is introduced to enable research on driver behavior understanding from untrimmed data sequences. As the first step, baseline algorithms for driver behavior detection are trained and tested to demonstrate the feasibility of the proposed task.

****

Learning Attribute Representations With Localization for Flexible Fashion Search

Kenan E. Ak, Ashraf A. Kassim, Joo Hwee Lim, Jo Yew Tham; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7708-7717

In this paper, we investigate ways of conducting a detailed fashion search using query images and attributes. A credible fashion search platform should be able to (1) find images that share the same attributes as the query image, (2) allow users to manipulate certain attributes, e.g. replace collar attribute from round to v-neck, and (3) handle region-specific attribute manipulations, e.g. replacing the color attribute of the sleeve region without changing the color attribute of other regions. A key challenge to be addressed is that fashion products have multiple attributes and it is important for each of these attributes to have representative features. To address these challenges, we propose the FashionSearchNet which uses a weakly supervised localization method to extract regions of attributes. By doing so, unrelated features can be ignored thus improving the similarity learning. Also, FashionSearchNet incorporates a new procedure that enables region awareness to be able to handle region-specific requests. FashionSearchNet outperforms the most recent fashion search techniques and is shown to be able to carry out different search scenarios using the dynamic queries.

****

Bidirectional Retrieval Made Simple

Jônatas Wehrmann, Rodrigo C. Barros; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7718-7726

This paper provides a very simple yet effective character-level architecture for learning bidirectional retrieval models. Aligning multimodal content is particularly challenging considering the difficulty in finding semantic correspondence between images and descriptions. We introduce an efficient character-level inception module, designed to learn textual semantic embeddings by convolving raw characters in distinct granularity levels. Our approach is capable of explicitly encoding hierarchical information from distinct base-level representations (e.g., characters, words, and sentences) into a shared multimodal space, where it maps the semantic correspondence between images and descriptions via a contrastive pairwise loss function that minimizes order-violations. Models generated by our approach are far more robust to input noise than state-of-the-art strategies based on word-embeddings. Despite being conceptually much simpler and requiring fewer parameters, our models outperform the state-of-the-art approaches by 4.8% in the task of description retrieval and 2.7% (absolute R@1 values) in the task of image retrieval in the popular MS COCO retrieval dataset. Finally, we show that our models present solid performance for text classification as well, specially in multilingual and noisy domains.

****

Learning Multi-Instance Enriched Image Representations via Non-Greedy Ratio Maximization of the l1-Norm Distances

Kai Liu, Hua Wang, Feiping Nie, Hao Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7727-7735

Multi-instance learning (MIL) has demonstrated its usefulness in many real-world image applications in recent years. However, two critical challenges prevent one from effectively using MIL in practice. First, existing MIL methods routinely model the predictive targets using the instances of input images, but rarely utilize an input image as a whole. As a result, the useful information conveyed by the holistic representation of an input image could be potentially lost. Second, the varied numbers of the instances of the input images in a data set make it infeasible to use traditional learning models that can only deal with single-vector inputs. To tackle these two challenges, in this paper we propose a novel image representation learning method that can integrate the local patches (the instances) of an input image (the bag) and its holistic representation into one single-vector representation. Our new method first learns a projection to preserve both global and local consistencies of the instances of an input image. It then projects the holistic representation of the same image into the learned subspace for information enrichment. Taking into account the content and characterization variations in natural scenes and photos, we develop an objective that maximizes the ratio of the summations of a number of L1-norm distances, which is difficult to solve in general. To solve our objective, we derive a new efficient non-greedy iterative algorithm and rigorously prove its convergence. Promising results in extensive experiments have demonstrated improved performances of our new method that validate its effectiveness.

****

Learning Visual Knowledge Memory Networks for Visual Question Answering

Zhou Su, Chen Zhu, Yinpeng Dong, Dongqi Cai, Yurong Chen, Jianguo Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7736-7745

Visual question answering (VQA) requires joint comprehension of images and natural language questions, where many questions can't be directly or clearly answered from visual content but require reasoning from structured human knowledge with confirmation from visual content. This paper proposes visual knowledge memory network (VKMN) to address this issue, which seamlessly incorporates structured human knowledge and deep visual features into memory networks in an end-to-end learning framework. Comparing to existing methods for leveraging external knowledge for supporting VQA, this paper stresses more on two missing mechanisms. First is the mechanism for integrating visual contents with knowledge facts. VKMN handles this issue by embedding knowledge triples (subject, relation, target) and deep visual features jointly into the visual knowledge features. Second is the mechanism for handling multiple knowledge facts expanding from question and answer pairs. VKMN stores joint embedding using key-value pair structure in the memory networks so that it is easy to handle multiple facts. Experiments show that the proposed method achieves promising results on both VQA v1.0 and v2.0 benchmarks, while outperforms state-of-the-art methods on the knowledge-reasoning related questions.

****

Visual Grounding via Accumulated Attention

Chaorui Deng, Qi Wu, Qingyao Wu, Fuyuan Hu, Fan Lyu, Mingkui Tan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7746-7755

Visual Grounding (VG) aims to locate the most relevant object or region in an im

age, based on a natural language query. The query can be a phrase, a sentence or even a multi-round dialogue. There are three main challenges in VG: 1) what is the main focus in a query; 2) how to understand an image; 3) how to locate an object. Most existing methods combine all the information curtly, which may suffer from the problem of information redundancy (i.e. ambiguous query, complicated image and a large number of objects). In this paper, we formulate these challenges as three attention problems and propose an accumulated attention (A-ATT) mechanism to reason among them jointly. Our A-ATT mechanism can circularly accumulate the attention for useful information in image, query, and objects, while the noises are ignored gradually. We evaluate the performance of A-ATT on four popular datasets (namely ReferCOCO, ReferCOCO+, ReferCOCOg, and Guesswhat?!), and the experimental results show the superiority of the proposed method in term of accuracy.
****
Beyond Trade-Off: Accelerate FCN-Based Face Detector With Higher Accuracy
Guanglu Song, Yu Liu, Ming Jiang, Yujie Wang, Junjie Yan, Biao Leng; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7756-7764
Fully convolutional neural network (FCN) has been dominating the game of face detection task for a few years with its congenital capability of sliding-window-searching with shared kernels, which boiled down all the redundant calculation, and most recent state-of-the-art methods such as Faster-RCNN, SSD, YOLO and FPN use FCN as their backbone. So here comes one question: Can we find a universal strategy to further accelerate FCN with higher accuracy, so could accelerate all the recent FCN-based methods? To analyze this, we decompose the face searching space into two orthogonal directions, `scale' and `spatial'. Only a few coordinates in the space expanded by the two base vectors indicate foreground. So if FCN could ignore most of the other points, the searching space and false alarm should be significantly boiled down. Based on this philosophy, a novel method named scale estimation and spatial attention proposal (S^2AP) is proposed to pay attention to some specific scales in image pyramid and valid locations in each scales layer. Furthermore, we adopt a masked convolution operation based on the attention result to accelerate FCN calculation. Experiments show that FCN-based method RPN can be accelerated by about 4X with the help of S^2AP and masked-FCN and at the same time it can also achieve the state-of-the-art on FDDB, AFW and MALF face detection benchmarks as well.
****
PackNet: Adding Multiple Tasks to a Single Network by Iterative Pruning
Arun Mallya, Svetlana Lazebnik; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7765-7773
This paper presents a method for adding multiple tasks to a single deep neural network while avoiding catastrophic forgetting. Inspired by network pruning techniques, we exploit redundancies in large deep networks to free up parameters that can then be employed to learn new tasks. By performing iterative pruning and network re-training, we are able to sequentially ``pack'' multiple tasks into a single network while ensuring minimal drop in performance and minimal storage overhead. Unlike prior work that uses proxy losses to maintain accuracy on older tasks, we always optimize for the task at hand. We perform extensive experiments on a variety of network architectures and large-scale datasets, and observe much better robustness against catastrophic forgetting than prior work. In particular, we are able to add three fine-grained classification tasks to a single ImageNet-trained VGG-16 network and achieve accuracies close to those of separately trained networks for each task.
****
Repulsion Loss: Detecting Pedestrians in a Crowd
Xinlong Wang, Tete Xiao, Yuning Jiang, Shuai Shao, Jian Sun, Chunhua Shen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7774-7783
Detecting individual pedestrians in a crowd remains a challenging problem since the pedestrians often gather together and occlude each other in real-world scena

rios. In this paper, we first explore how a state-of-the-art pedestrian detector is harmed by crowd occlusion via experimentation, providing insights into the crowd occlusion problem. Then, we propose a novel bounding box regression loss specifically designed for crowd scenes, termed repulsion loss. This loss is driven by two motivations: the attraction by target, and the repulsion by other surrounding objects. The repulsion term prevents the proposal from shifting to surrounding objects thus leading to more crowd-robust localization. Our detector trained by repulsion loss outperforms the state-of-the-art methods with a significant improvement in occlusion cases.

****

Neural Sign Language Translation

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, Richard Bowden; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7784-7793

Sign Language Recognition (SLR) has been an active research field for the last two decades. However, most research to date has considered SLR as a naive gesture recognition problem. SLR seeks to recognize a sequence of continuous signs but neglects the underlying rich grammatical and linguistic structures of sign language that differ from spoken language. In contrast, we introduce the Sign Language Translation (SLT) problem. Here, the objective is to generate spoken language translations from sign language videos, taking into account the different word orders and grammar. We formalize SLT in the framework of Neural Machine Translation (NMT) for both end-to-end and pretrained settings (using expert knowledge). This allows us to jointly learn the spatial representations, the underlying language model, and the mapping between sign and spoken language. To evaluate the performance of Neural SLT, we collected the first publicly available Continuous SLT dataset, RWTH-PHOENIX-Weather 2014T. It provides spoken language translations and gloss level annotations for German Sign Language videos of weather broadcasts. Our dataset contains over .95M frames with >67K signs from a sign vocabulary of >1K and >99K words from a German vocabulary of >2.8K. We report quantitative and qualitative results for various SLT setups to underpin future research in this newly established field. The upper bound for translation performance is calculated at 19.26 BLEU-4, while our end-to-end frame-level and gloss-level tokenization networks were able to achieve 9.58 and 18.13 respectively.

****

Non-Local Neural Networks

Xiaolong Wang, Ross Girshick, Abhinav Gupta, Kaiming He; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7794-7803

Both convolutional and recurrent operations are building blocks that process one local neighborhood at a time. In this paper, we present non-local operations as a generic family of building blocks for capturing long-range dependencies. Inspired by the classical non-local means method in computer vision, our non-local operation computes the response at a position as a weighted sum of the features at all positions. This building block can be plugged into many computer vision architectures. On the task of video classification, even without any bells and whistles, our non-local models can compete or outperform current competition winners on both Kinetics and Charades datasets. In static image recognition, our non-local models improve object detection/segmentation and pose estimation on the COCO suite of tasks. Code will be made available.

****

LAMV: Learning to Align and Match Videos With Kernelized Temporal Layers

Lorenzo Baraldi, Matthijs Douze, Rita Cucchiara, Hervé Jégou; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7804-7813

This paper considers a learnable approach for comparing and aligning videos. Our architecture builds upon and revisits temporal match kernels within neural networks: we propose a new temporal layer that finds temporal alignments by maximizing the scores between two sequences of vectors, according to a time-sensitive similarity metric parametrized in the Fourier domain. We learn this layer with a t emporal proposal strategy, in which we minimize a triplet loss that takes into account both the localization accuracy and the recognition rate. We evaluate our approach on video alignment, copy detection and event retrieval. Our approach outperforms the state on the art on temporal video alignment and video copy detection datasets in comparable setups. It also attains the best reported results for particular event search, while precisely aligning videos.

****

Optimizing Video Object Detection via a Scale-Time Lattice

Kai Chen, Jiaqi Wang, Shuo Yang, Xingcheng Zhang, Yuanjun Xiong, Chen Change Loy, Dahua Lin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7814-7823

High-performance object detection relies on expensive convolutional networks to compute features, often leading to significant challenges in applications, e.g. those that re- quire detecting objects from video streams in real time. The key to this problem is to trade accuracy for efficiency in an effective way, i.e. reducing the computing cost while maintaining competitive performance. To seek a good balance, previous efforts usually focus on optimizing the model architectures. This paper explores an alternative approach, that is, to reallocate the computation over a scale-time space. The basic idea is to perform expensive detection sparsely and propagate the results across both scales and time with substantially cheaper networks, by exploiting the strong correlations among them. Specifically, we present a unified framework that integrates detection, temporal propagation, and across-scale refinement on a Scale-Time Lattice. On this framework, one can explore various strategies to balance performance and cost. Taking advantage of this flexibility, we further develop an adaptive scheme with the detector invoked on demand and thus obtain improved tradeoff. On ImageNet VID dataset, the proposed method can achieve a competitive mAP 79.6% at 20 fps, or 79.0% at 62 fps as a performance/speed tradeoff.

****

Learning Compressible 360° Video Isomers

Yu-Chuan Su, Kristen Grauman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7824-7833

Standard video encoders developed for conventional narrow field-of-view video are widely applied to 360° video as well, with reasonable results. However, while this approach commits arbitrarily to a projection of the spherical frames, we observe that some orientations of a 360° video, once projected, are more compressible than others. We introduce an approach to predict the sphere rotation that will yield the maximal compression rate. Given video clips in their original encoding, a convolutional neural network learns the association between a clip’s visual content and its compressibility at different rotations of a cubemap projection. Given a novel video, our learning-based approach efficiently infers the most compressible direction in one shot, without repeated rendering and compression of the source video. We validate our idea on thousands of video clips and multiple popular video codecs. The results show that this untapped dimension of 360° compression has substantial potential—“good” rotations are typically 8–10% more compressible than bad ones, and our learning approach can predict them reliably 82% of the time.

****

Attention Clusters: Purely Attention Based Local Feature Integration for Video Classification

Xiang Long, Chuang Gan, Gerard de Melo, Jiajun Wu, Xiao Liu, Shilei Wen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7834-7843

Recently, substantial research effort has focused on how to apply CNNs or RNNs to better capture temporal patterns in videos, so as to improve the accuracy of video classification. In this paper, however, we show that temporal information, especially longer-term patterns, may not be necessary to achieve competitive results on common trimmed video classification datasets. We investigate the potential of a purely attention based local feature integration. Accounting for the characteristics of such features in video classification, we propose a local featur e integration framework based on attention clusters, and introduce a shifting operation to capture more diverse signals. We carefully analyze and compare the effect of different attention mechanisms, cluster sizes, and the use of the shifting operation, and also investigate the combination of attention clusters for multimodal integration. We demonstrate the effectiveness of our framework on three real-world video classification datasets. Our model achieves competitive results across all of these. In particular, on the large-scale Kinetics dataset, our framework obtains an excellent single model accuracy of 79.4% in terms of the top-1 and 94.0% in terms of the top-5 accuracy on the validation set.

****

What Have We Learned From Deep Representations for Action Recognition?

Christoph Feichtenhofer, Axel Pinz, Richard P. Wildes, Andrew Zisserman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7844-7853

As the success of deep models has led to their deployment in all areas of computer vision, it is increasingly important to understand how these representations work and what they are capturing. In this paper, we shed light on deep spatiotemporal representations by visualizing what two-stream models have learned in order to recognize actions in video. We show that local detectors for appearance and motion objects arise to form distributed representations for recognizing human actions. Key observations include the following. First, cross-stream fusion enables the learning of true spatiotemporal features rather than simply separate appearance and motion features. Second, the networks can learn local representations that are highly class specific, but also generic representations that can serve a range of classes. Third, throughout the hierarchy of the network, features become more abstract and show increasing invariance to aspects of the data that are unimportant to desired distinctions (e.g. motion patterns across various speeds). Fourth, visualizations can be used not only to shed light on learned representations, but also to reveal idiosyncracies of training data and to explain failure cases of the system.

****

Controllable Video Generation With Sparse Trajectories

Zekun Hao, Xun Huang, Serge Belongie; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7854-7863

Video generation and manipulation is an important yet challenging task in computer vision. Existing methods usually lack ways to explicitly control the synthesized motion. In this work, we present a conditional video generation model that allows detailed control over the motion of the generated video. Given the first frame and sparse motion trajectories specified by users, our model can synthesize a video with corresponding appearance and motion. We propose to combine the advantage of copying pixels from the given frame and hallucinating the lightness difference from scratch which help generate sharp video while keeping the model robust to occlusion and lightness change. We also propose a training paradigm that calculate trajectories from video clips, which eliminated the need of annotated training data. Experiments on several standard benchmarks demonstrate that our approach can generate realistic videos comparable to state-of-the-art video generation and video prediction methods while the motion of the generated videos can correspond well with user input.

****

Representing and Learning High Dimensional Data With the Optimal Transport Map From a Probabilistic Viewpoint

Serim Park, Matthew Thorpe; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7864-7872

In this paper, we propose a generative model in the space of diffeomorphic deformation maps. More precisely, we utilize the Kantarovich-Wasserstein metric and accompanying geometry to represent an image as a deformation from templates. Moreover, we incorporate a probabilistic viewpoint by assuming that each image is locally generated from a reference image. We capture the local structure by modelling the tangent planes at reference images. %; we assume that each image is generated from one of finite number of tangent planes. % by an unobserved discrete r andom variable that indexes the tangent plane the image belongs to.   Once basis vectors for each tangent plane are learned via probabilistic PCA, we can sample a local coordinate, that can be inverted back to image space exactly. With experiments using 4 different datasets, we show that the generative tangent plane model in the optimal transport (OT) manifold can be learned with small numbers of images and can be used to create infinitely many `unseen' images. In addition, the Bayesian classification accompanied with the probabilist modeling of the tangent planes shows improved accuracy over that done in the image space. Combining the results of our experiments supports our claim that certain datasets can be better represented with the Kantarovich-Wasserstein metric. We envision that the proposed method could be a practical solution to learning and representing data that is generated with templates in situatons where only limited numbers of data points are available.

****

CLIP-Q: Deep Network Compression Learning by In-Parallel Pruning-Quantization

Frederick Tung, Greg Mori; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7873-7882

Deep neural networks enable state-of-the-art accuracy on visual recognition tasks such as image classification and object detection. However, modern deep networks contain millions of learned weights; a more efficient utilization of computation resources would assist in a variety of deployment scenarios, from embedded platforms with resource constraints to computing clusters running ensembles of networks. In this paper, we combine network pruning and weight quantization in a single learning framework that performs pruning and quantization jointly, and in parallel with fine-tuning. This allows us to take advantage of the complementary nature of pruning and quantization and to recover from premature pruning errors, which is not possible with current two-stage approaches. Our proposed CLIP-Q method (Compression Learning by In-Parallel Pruning-Quantization) compresses AlexNet by 51-fold, GoogLeNet by 10-fold, and ResNet-50 by 15-fold, while preserving the uncompressed network accuracies on ImageNet.

****

Inference in Higher Order MRF-MAP Problems With Small and Large Cliques

Ishant Shanu, Chetan Arora, S.N. Maheshwari; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7883-7891

Higher Order MRF-MAP formulation has been a popular technique for solving many problems in computer vision. Inference in a general MRF-MAP problem is NP Hard, but can be performed in polynomial time for the special case when potential functions are submodular. Two popular combinatorial approaches for solving such formulations are flow based and polyhedral approaches. Flow based approaches work well with small cliques and in that mode can handle problems with millions of variables. Polyhedral approaches can handle large cliques but in small numbers. We show in this paper that the variables in these seemingly disparate techniques can be mapped to each other. This allows us to combine the two styles in a joint framework exploiting the strength of both of them. Using the proposed joint framework, we are able to perform tractable inference in MRF-MAP problems with millions of variables and a mix of small and large cliques, a formulation which can not be solved by either of the two styles individually. We show applicability of this hybrid framework on object segmentation problem as an example of a situation where quality of results is significantly better than systems which are based only on the use of small or large cliques.

****

ROAD: Reality Oriented Adaptation for Semantic Segmentation of Urban Scenes

Yuhua Chen, Wen Li, Luc Van Gool; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7892-7901

Exploiting synthetic data to learn deep models has attracted increasing attention in recent years. However, the intrinsic domain difference between synthetic and real images usually causes a significant performance drop when applying the learned model to real world scenarios. This is mainly due to two reasons: 1) the model overfits to synthetic images, making the convolutional filters incompetent to extract informative representation for real images; 2) there is a distributio n difference between synthetic and real data, which is also known as the domain adaptation problem. To this end, we propose a new reality oriented adaptation approach for urban scene semantic segmentation by learning from synthetic data. First, we propose a target guided distillation approach to learn the real image style, which is achieved by training the segmentation model to imitate a pretrained real style model using real images. Second, we further take advantage of the intrinsic spatial structure presented in urban scene images, and propose a spatial-aware adaptation scheme to effectively align the distribution of two domains. These two modules can be readily integrated with existing state-of-the-art semantic segmentation networks to improve their generalizability when adapting from synthetic to real urban scenes. We evaluate the proposed method on Cityscapes dataset by adapting from GTAV and SYNTHIA datasets, where the results demonstrate the effectiveness of our method.

****

Eye In-Painting With Exemplar Generative Adversarial Networks

Brian Dolhansky, Cristian Canton Ferrer; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7902-7911

This paper introduces a novel approach to in-painting where the identity of the object to remove or change is preserved and accounted for at inference time: Exemplar GANs (ExGANs). ExGANs are a type of conditional GAN that utilize exemplar information to produce high-quality, personalized in-painting results. We propose using exemplar information in the form of a reference image of the region to in-paint, or a perceptual code describing that object. Unlike previous conditional GAN formulations, this extra information can be inserted at multiple points within the adversarial network, thus increasing its descriptive power. We show that ExGANs can produce photo-realistic personalized in-painting results that are both perceptually and semantically plausible by applying them to the task of closed-to-open eye in-painting in natural pictures. A new benchmark dataset is also introduced for the task of eye in-painting for future comparisons.

****

ClcNet: Improving the Efficiency of Convolutional Neural Network Using Channel Local Convolutions

Dong-Qing Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7912-7919

Depthwise convolution and grouped convolution has been successfully applied to improve the efficiency of convolutional neural network (CNN). We suggest that these models can be considered as special cases of a generalized convolution operation, named channel local convolution(CLC), where an output channel is computed using a subset of the input channels. This definition entails computation dependency relations between input and output channels, which can be represented by a channel dependency graph(CDG). By modifying the CDG of grouped convolution, a new CLC kernel named interlaced grouped convolution (IGC) is created. Stacking IGC and GC kernels results in a convolution block (named CLC Block) for approximating regular convolution. By resorting to the CDG as an analysis tool, we derive the rule for setting the meta-parameters of IGC and GC and the framework for minimizing the computational cost. A new CNN model named clcNet is then constructed using CLC blocks, which shows significantly higher computational efficiency and fewer parameters compared to state-of-the-art networks, when being tested using the ImageNet-1K dataset.

****

Towards Effective Low-Bitwidth Convolutional Neural Networks

Bohan Zhuang, Chunhua Shen, Mingkui Tan, Lingqiao Liu, Ian Reid; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7920-7928

This paper tackles the problem of training a deep convolutional neural network with both low-precision weights and low-bitwidth activations. Optimizing a low-precision network is very challenging since the training process can easily get trapped in a poor local minima, which results in substantial accuracy loss. To mitigate this problem, we propose three simple-yet-effective approaches to improve the network training. First, we propose to use a two-stage optimization strateg y to progressively find good local minima. Specifically, we propose to first optimize a net with quantized weights and then quantized activations. This is in contrast to the traditional methods which optimize them simultaneously. Second, following a similar spirit of the first method, we propose another progressive optimization approach which progressively decreases the bit-width from high-precision to low-precision during the course of training. Third, we adopt a novel learning scheme to jointly train a full-precision model alongside the low-precision one. By doing so, the full-precision model provides hints to guide the low-precision model training. Extensive experiments on various datasets (ie, CIFAR-100 and ImageNet) show the effectiveness of the proposed methods. To highlight, using our methods to train a 4-bit precision network leads to no performance decrease in comparison with its full-precision counterpart with standard network architectures (ie, AlexNet and ResNet-50).

****

Stochastic Downsampling for Cost-Adjustable Inference and Improved Regularization in Convolutional Networks

Jason Kuen, Xiangfei Kong, Zhe Lin, Gang Wang, Jianxiong Yin, Simon See, Yap-Peng Tan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7929-7938

It is desirable to train convolutional networks (CNNs) to run more efficiently during inference. In many cases however, the computational budget that the system has for inference cannot be known beforehand during training, or the inference budget is dependent on the changing real-time resource availability. Thus, it is inadequate to train just inference-efficient CNNs, whose inference costs are not adjustable and cannot adapt to varied inference budgets. We propose a novel approach for cost-adjustable inference in CNNs - Stochastic Downsampling Point (SDPoint). During training, SDPoint applies feature map downsampling to a random point in the layer hierarchy, with a random downsampling ratio. The different stochastic downsampling configurations known as SDPoint instances (of the same model) have computational costs different from each other, while being trained to minimize the same prediction loss. Sharing network parameters across different instances provides significant regularization boost. During inference, one may handpick a SDPoint instance that best fits the inference budget. The effectiveness of SDPoint, as both a cost-adjustable inference approach and a regularizer, is validated through extensive experiments on image classification.

****

Face Aging With Identity-Preserved Conditional Generative Adversarial Networks

Zongwei Wang, Xu Tang, Weixin Luo, Shenghua Gao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7939-7947

Face aging is of great importance for cross-age recognition and entertainment related applications. However, the lack of labeled faces of the same person across a long age range makes it challenging. Because of different aging speed of different persons, our face aging approach aims at synthesizing a face whose target age lies in some given age group instead of synthesizing a face with a certain age. By grouping faces with target age together, the objective of face aging is equivalent to transferring aging patterns of faces within the target age group to the face whose aged face is to be synthesized. Meanwhile, the synthesized face should have the same identity with the input face. Thus we propose an Identity-Preserved Conditional Generative Adversarial Networks (IPCGANs) framework, in which a Conditional Generative Adversarial Networks module functions as generating a face that looks realistic and is with the target age, an identity-preserved module preserves the identity information and an age classifier forces the generated face with the target age. Both qualitative and quantitative experiments show that our method can generate more realistic faces in terms of image quality, person identity and age consistency with human observations.

****

Unsupervised Cross-Dataset Person Re-Identification by Transfer Learning of Spatial-Temporal Patterns

Jianming Lv, Weihang Chen, Qing Li, Can Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7948-7956

Most of the proposed person re-identification algorithms conduct supervised training and testing on single labeled datasets with small size, so directly deploying these trained models to a large-scale real-world camera network may lead to poor performance due to underfitting. It is challenging to incrementally optimize the models by using the abundant unlabeled data collected from the target domain. To address this challenge, we propose an unsupervised incremental learning algorithm, TFusion, which is aided by the transfer learning of the pedestrians' spatio-temporal patterns in the target domain. Specifically, the algorithm firstly transfers the visual classifier trained from small labeled source dataset to the unlabeled target dataset so as to learn the pedestrians' spatial-temporal patterns. Secondly, a Bayesian fusion model is proposed to combine the learned spatio-temporal patterns with visual features to achieve a significantly improved classifier. Finally, we propose a learning-to-rank based mutual promotion procedure to incrementally optimize the classifiers based on the unlabeled data in the target domain. Comprehensive experiments based on multiple real surveillance datasets are conducted, and the results show that our algorithm gains significant improvement compared with the state-of-art cross-dataset unsupervised person re-identification algorithms.

****

Feature Quantization for Defending Against Distortion of Images

Zhun Sun, Mete Ozay, Yan Zhang, Xing Liu, Takayuki Okatani; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7957-7966

In this work, we address the problem of improving robustness of convolutional neural networks (CNNs) to image distortion. We argue that higher moment statistics of feature distributions can be shifted due to image distortion, and the shift leads to performance decrease and cannot be reduced by ordinary normalization methods as observed in our experimental analyses. In order to mitigate this effect, we propose an approach base on feature quantization. To be specific, we propose to employ three different types of additional non-linearity in CNNs: i) a floor function with scalable resolution, ii) a power function with learnable exponents, and iii) a power function with data-dependent exponents. In the experiments, we observe that CNNs which employ the proposed methods obtain better performance in both generalization performance and robustness for various distortion types for large scale benchmark datasets. For instance, a ResNet-50 model equipped with the proposed method (+HPOW) obtains 6.95%, 5.26% and 5.61% better accuracy on the ILSVRC-12 classification tasks using images distorted with motion blur, salt and pepper and mixed distortions.

****

Tagging Like Humans: Diverse and Distinct Image Annotation

Baoyuan Wu, Weidong Chen, Peng Sun, Wei Liu, Bernard Ghanem, Siwei Lyu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7967-7975

In this work we propose a new automatic image annotation model, dubbed diverse and distinct image annotation (D2IA). The generative model D2IA is inspired by the ensemble of human annotations, which create semantically relevant, yet distinct and diverse tags. In D2IA, we generate a relevant and distinct tag subset, in which the tags are relevant to the image contents and semantically distinct to each other, using sequential sampling from a determinantal point process (DPP) model. Multiple such tag subsets that cover diverse semantic aspects or diverse semantic levels of the image contents are generated by randomly perturbing the DPP sampling process. We leverage a generative adversarial network (GAN) model to train D2IA. We perform extensive experiments including quantitative and qualitative comparisons, as well as human subject studies, on two benchmark datasets to demonstrate that the proposed model can produce more diverse and distinct tags than the state-of-the-arts.

****

Re-Weighted Adversarial Adaptation Network for Unsupervised Domain Adaptation

Qingchao Chen, Yang Liu, Zhaowen Wang, Ian Wassell, Kevin Chetty; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp

. 7976-7985

Unsupervised Domain Adaptation (UDA) aims to transfer domain knowledge from existing well-defined tasks to new ones where labels are unavailable. In the real-world applications, as the domain (task) discrepancies are usually uncontrollable, it is significantly motivated to match the feature distributions even if the domain discrepancies are disparate. Additionally, as no label is available in the target domain, how to successfully adapt the classifier from the source to the target domain still remains an open question. In this paper, we propose the Re-weighted Adversarial Adaptation Network (RAAN) to reduce the feature distribution divergence and adapt the classifier when domain discrepancies are disparate. Specifically, to alleviate the need of common supports in matching the feature distribution, we choose to minimize optimal transport (OT) based Earth-Mover (EM) distance and reformulate it to a minimax objective function. Utilizing this, RAAN can be trained in an end-to-end and adversarial manner. To further adapt the classifier, we propose to match the label distribution and embed it into the adversarial training. Finally, after extensive evaluation of our method using UDA datasets of varying difficulty, RAAN achieved the state-of-the-art results and outperformed other methods by a large margin when the domain shifts are disparate.

****

Inferring Semantic Layout for Hierarchical Text-to-Image Synthesis

Seunghoon Hong, Dingdong Yang, Jongwook Choi, Honglak Lee; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7986-7994

We propose a novel hierarchical approach for text-to-image synthesis by inferring semantic layout. Instead of learning a direct mapping from text to image, our algorithm decomposes the generation process into multiple steps, in which it first constructs a semantic layout from the text by the layout generator and converts the layout to an image by the image generator. The proposed layout generator progressively constructs a semantic layout in a coarse-to-fine manner by generating object bounding boxes and refining each box by estimating object shapes inside the box. The image generator synthesizes an image conditioned on the inferred semantic layout, which provides a useful semantic structure of an image matching with the text description. Our model not only generates semantically more meaningful images, but also allows automatic annotation of generated images and user-controlled generation process by modifying the generated scene layout. We demonstrate the capability of the proposed model on challenging MS-COCO dataset and show that the model can substantially improve the image quality, interpretability of output and semantic alignment to input text over existing approaches.

****

Regularizing RNNs for Caption Generation by Reconstructing the Past With the Present

Xinpeng Chen, Lin Ma, Wenhao Jiang, Jian Yao, Wei Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7995-8003

Recently, caption generation with an encoder-decoder framework has been extensively studied and applied in different domains, such as image captioning, code captioning, and so on. In this paper, we propose a novel architecture, namely Auto-Reconstructor Network (ARNet), which, coupling with the conventional encoder-decoder framework, works in an end-to-end fashion to generate captions. ARNet aims at reconstructing the previous hidden state with the present one, besides behaving as the input-dependent transition operator. Therefore, ARNet encourages the current hidden state to embed more information from the previous one, which can help regularize the transition dynamics of recurrent neural networks (RNNs). Extensive experimental results show that our proposed ARNet boosts the performance over the existing encoder-decoder models on both image captioning and source code captioning tasks. Additionally, ARNet remarkably reduces the discrepancy between training and inference processes for caption generation. Furthermore, the performance on permuted sequential MNIST demonstrates that ARNet can effectively regularize RNN, especially on modeling long-term dependencies. Our code is available at: https://github.com/chenxinpeng/ARNet.

****

Unsupervised Domain Adaptation With Similarity Learning

Pedro O. Pinheiro; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8004-8013

The objective of unsupervised domain adaptation is to leverage features from a labeled source domain and learn a classifier for an unlabeled target domain, with a similar but different data distribution. Most deep learning approaches consist of two steps: (i) learn features that preserve a low risk on labeled samples (source domain) and (ii) make the features from both domains to be as indistinguishable as possible, so that a classifier trained on the source can also be applied on the target domain. In general, the classifiers in step (i) consist of fully-connected layers applied directly on the indistinguishable features learned in (ii). In this paper, we propose a different way to do the classification, using similarity learning. The proposed method learns a pairwise similarity function in which classification can be performed by computing distances between prototype representations of each category. The domain-invariant features and the categorical prototype representations are learned jointly and in an end-to-end fashion. At inference time, images from the target domain are compared to the prototypes and the label associated with the one that best matches the image is outputed. The approach is simple, scalable and effective. We show that our model achieves state-of-the-art performance in different large-scale unsupervised domain adaptation scenarios.

****

Learning Deep Sketch Abstraction

Umar Riaz Muhammad, Yongxin Yang, Yi-Zhe Song, Tao Xiang, Timothy M. Hospedales; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8014-8023

Human free-hand sketches have been studied in various contexts including sketch recognition, synthesis and fine-grained sketch-based image retrieval (FG-SBIR). A fundamental challenge for sketch analysis is to deal with drastically different human drawing styles, particularly in terms of abstraction level. In this work, we propose the first stroke-level sketch abstraction model based on the insight of sketch abstraction as a process of trading off between the recognizability of a sketch and the number of strokes used to draw it. Concretely, we train a model for abstract sketch generation through reinforcement learning of a stroke removal policy that learns to predict which strokes can be safely removed without affecting recognizability. We show that our abstraction model can be used for various sketch analysis tasks including: (1) modeling stroke saliency and understanding the decision of sketch recognition models, (2) synthesizing sketches of variable abstraction for a given category, or reference object instance in a photo, and (3) training a FG-SBIR model with photos only, bypassing the expensive photo-sketch pair collection step.

****

Matching Adversarial Networks

Gellért Máttyus, Raquel Urtasun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8024-8032

Generative Adversarial Nets (GANs) and Conditonal GANs (CGANs) show that using a trained network as loss function (discriminator) enables to synthesize highly structured outputs (e.g. natural images). However, applying a discriminator network as a universal loss function for common supervised tasks (e.g. semantic segmentation, line detection, depth estimation) is considerably less successful. We argue that the main difficulty of applying CGANs to supervised tasks is that the generator training consists of optimizing a loss function that does not depend directly on the ground truth labels.   To overcome this, we propose to replace the discriminator with a matching network taking into account both the ground truth outputs as well as the generated examples. As a consequence, the generator loss function also depends on the targets of the training examples, thus facilitating learning. We demonstrate on three computer vision tasks that this approach can significantly outperform CGANs achieving comparable or superior results to task-specific solutions and results in stable training. Importantly, this is a general approach that does not require the use of task-specific loss functions.

****

SoS-RSC: A Sum-of-Squares Polynomial Approach to Robustifying Subspace Clustering Algorithms

Mario Sznaier, Octavia Camps; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8033-8041

This paper addresses the problem of subspace clustering in the presence of outliers. Typically, this scenario is handled through a regularized optimization, whose computational complexity scales polynomially with the size of the data. Further, the regularization terms need to be manually tuned to achieve optimal performance. To circumvent these difficulties, in this paper we propose an outlier removal algorithm based on evaluating a suitable sum-ofsquares polynomial, computed directly from the data. This algorithm only requires performing two singular value decompositions of fixed size, and provides certificates on the probability of misclassifying outliers as inliers.

****

Resource Aware Person Re-Identification Across Multiple Resolutions

Yan Wang, Lequn Wang, Yurong You, Xu Zou, Vincent Chen, Serena Li, Gao Huang, Bharath Hariharan, Kilian Q. Weinberger; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8042-8051

Not all people are equally easy to identify: color statistics might be enough for some cases while others might require careful reasoning about high- and low-level details. However, prevailing person re-identification(re-ID) methods use one-size-fits-all high-level embeddings from deep convolutional networks for all cases. This might limit their accuracy on difficult examples or makes them needlessly expensive for the easy ones. To remedy this, we present a new person re-ID model that combines effective embeddings built on multiple convolutional network layers, trained with deep-supervision. On traditional re-ID benchmarks, our method improves substantially over the previous state-of-the-art results on all five datasets that we evaluate on. We then propose two new formulations of the person re-ID problem under resource-constraints, and show how our model can be used to effectively trade off accuracy and computation in the presence of resource constraints.

****

Learning and Using the Arrow of Time

Donglai Wei, Joseph J. Lim, Andrew Zisserman, William T. Freeman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8052-8060

We seek to understand the arrow of time in videos -- what makes videos look like they are playing forwards or backwards? Can we visualize the cues? Can the arrow of time be a supervisory signal useful for activity analysis? To this end, we build three large-scale video datasets and apply a learning-based approach to these tasks.   To learn the arrow of time efficiently and reliably, we design a ConvNet suitable for extended temporal footprints and for class activation visualization, and study the effect of artificial cues, such as cinematographic conventions, on learning. Our trained model achieves state-of-the-art performance on large-scale real-world video datasets. Through cluster analysis and localization of important regions for the prediction, we examine learned visual cues that are consistent among many samples and show when and where they occur. Lastly, we use the trained ConvNet for two applications: self-supervision for action recognition, and video forensics -- determining whether Hollywood film clips have been deliberately reversed in time, often used as special effects.

****

Neural Style Transfer via Meta Networks

Falong Shen, Shuicheng Yan, Gang Zeng; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8061-8069

In this paper we propose a noval method to generate the specified network parameters through one feed-forward propagation in the meta networks for neural style transfer. Recent works on style transfer typically need to train image transformation networks for every new style, and the style is encoded in the network parameters by enormous iterations of stochastic gradient descent, which lacks the ge

neralization ability to new style in the inference stage. To tackle these issues, we build a meta network which takes in the style image and generates a corresponding image transformation network directly. Compared with optimization-based methods for every style, our meta networks can handle an arbitrary new style within 19 milliseconds on one modern GPU card. The fast image transformation network generated by our meta network is only 449 KB, which is capable of real-time running on a mobile device. We also investigate the manifold of the style transfer networks by operating the hidden features from meta networks. Experiments have well validated the effectiveness of our method. Code and trained models will be released.
****
People, Penguins and Petri Dishes: Adapting Object Counting Models to New Visual Domains and Object Types Without Forgetting
Mark Marsden, Kevin McGuinness, Suzanne Little, Ciara E. Keogh, Noel E. O'Connor; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8070-8079
In this paper we propose a technique to adapt a convolutional neural network (CNN) based object counter to additional visual domains and object types while still preserving the original counting function. Domain-specific normalisation and scaling operators are trained to allow the model to adjust to the statistical distributions of the various visual domains. The developed adaptation technique is used to produce a singular patch-based counting regressor capable of counting various object types including people, vehicles, cell nuclei and wildlife. As part of this study a challenging new cell counting dataset in the context of tissue culture and patient diagnosis is constructed. This new collection, referred to as the Dublin Cell Counting (DCC) dataset, is the first of its kind to be made available to the wider computer vision community. State-of-the-art object counting performance is achieved in both the Shanghaitech (parts A and B) and Penguins datasets while competitive performance is observed on the TRANCOS and Modified Bone Marrow (MBM) datasets, all using a shared counting model.
****
HydraNets: Specialized Dynamic Architectures for Efficient Inference
Ravi Teja Mullapudi, William R. Mark, Noam Shazeer, Kayvon Fatahalian; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8080-8089
There is growing interest in improving the design of deep network architectures to be both accurate and low cost. This paper explores semantic specialization as a mechanism for improving the computational efficiency (accuracy-per-unit-cost) of inference in the context of image classification. Specifically, we propose a network architecture template called HydraNet, which enables state-of-the-art architectures for image classification to be transformed into dynamic architectures which exploit conditional execution for efficient inference. HydraNets are wide networks containing distinct components specialized to compute features for visually similar classes, but they retain efficiency by dynamically selecting only a small number of components to evaluate for any one input image. This design is made possible by a soft gating mechanism that encourages component specialization during training and accurately performs component selection during inference. We evaluate the HydraNet approach on both the CIFAR-100 and ImageNet classification tasks. On CIFAR, applying the HydraNet template to the ResNet and DenseNet family of models reduces inference cost by 2-4x while retaining the accuracy of the baseline architectures. On ImageNet, applying the HydraNet template improves accuracy up to 2.5% when compared to an efficient baseline architecture with similar inference cost.
****
SketchMate: Deep Hashing for Million-Scale Human Sketch Retrieval
Peng Xu, Yongye Huang, Tongtong Yuan, Kaiyue Pang, Yi-Zhe Song, Tao Xiang, Timothy M. Hospedales, Zhanyu Ma, Jun Guo; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8090-8098
We propose a deep hashing framework for sketch retrieval that, for the first time, works on a multi-million scale human sketch dataset.Leveraging on this large

dataset, we explore a few sketch-specific traits that were otherwise under-studied in prior literature. Instead of following the conventional sketch recognition task, we introduce the novel problem of sketch hashing retrieval which is not only more challenging, but also offers a better testbed for large-scale sketch analysis, since: (i) more fine-grained sketch feature learning is required to accommodate the large variations in style and abstraction, and (ii) a compact binary code needs to be learned at the same time to enable efficient retrieval.Key to our network design is the embedding of unique characteristics of human sketch, where (i) a two-branch CNN-RNN architecture is adapted to explore the temporal ordering of strokes, and (ii) a novel hashing loss is specifically designed to accommodate both the temporal and abstract traits of sketches. By working with a 3.8M sketch dataset,we show that state-of-the-art hashing models specifically engineered for static images fail to perform well on temporal sketch data. Our network on the other hand not only offers the best retrieval performance on various code sizes, but also yields the best generalization performance under a zero-shot setting and when re-purposed for sketch recognition.Such superior performances effectively demonstrate the benefit of our sketch-specific design.

****

From Source to Target and Back: Symmetric Bi-Directional Adaptive GAN

Paolo Russo, Fabio M. Carlucci, Tatiana Tommasi, Barbara Caputo; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8099-8108

The effectiveness of GANs in producing images according to a specific visual domain has shown potential in unsupervised domain adaptation. Source labeled images have been modified to mimic target samples for training classifiers in the target domain, and inverse mappings from the target to the source domain have also been evaluated, without new image generation. In this paper we aim at getting the best of both worlds by introducing a symmetric mapping among domains. We jointly optimize bi-directional image transformations combining them with target self-labeling. We define a new class consistency loss that aligns the generators in the two directions, imposing to preserve the class identity of an image passing through both domain mappings. A detailed analysis of the reconstructed images, a thorough ablation study and extensive experiments on six different settings confirm the power of our approach.

****

OLÉ: Orthogonal Low-Rank Embedding - A Plug and Play Geometric Loss for Deep Learning

José Lezama, Qiang Qiu, Pablo Musé, Guillermo Sapiro; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8109-8118

Deep neural networks trained using a softmax layer at the top and the cross-entropy loss are ubiquitous tools for image classification. Yet, this does not naturally enforce intra-class similarity nor inter-class margin of the learned deep representations. To simultaneously achieve these two goals, different solutions have been proposed in the literature, such as the pairwise or triplet losses. However, these carry the extra task of selecting pairs or triplets, and the extra computational burden of computing and learning for many combinations of them. In this paper, we propose a plug-and-play loss term for deep networks that explicitly reduces intra-class variance and enforces inter-class margin simultaneously, in a simple and elegant geometric manner. For each class, the deep features are collapsed into a learned linear subspace, or union of them, and inter-class subspaces are pushed to be as orthogonal as possible. Our proposed Orthogonal Low-rank Embedding (OLE) does not require carefully crafting pairs or triplets of samples for training, and works standalone as a classification loss, being the first reported deep metric learning framework of its kind. Because of the improved margin between features of different classes, the resulting deep networks generalize better, are more discriminative, and more robust. We demonstrate improved classification performance in general object recognition, plugging the proposed loss term into existing off-the-shelf architectures. In particular, we show the advantage of the proposed loss in the small data/model scenario, and we significantly advance the state-of-the-art on the Stanford STL-10 benchmark.

****

Efficient Parametrization of Multi-Domain Deep Neural Networks

Sylvestre-Alvise Rebuffi, Hakan Bilen, Andrea Vedaldi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8119-8127

A practical limitation of deep neural networks is their high degree of specialization to a single task and visual domain. In complex applications such as mobile platforms, this requires juggling several large models with detrimental effect on speed and battery life. Recently, inspired by the successes of transfer learning, several authors have proposed to learn instead universal, fixed feature extractors that, used as the first stage of any deep network, work well for all tasks and domains simultaneously. Nevertheless, such universal features are still somewhat inferior to specialized networks. To overcome this limitation, in this paper we propose to consider instead universal parametric families of neural networks, which still contain specialized problem-specific models, but that differ only by a small number of parameters. We study different designs for such parametrizations, including series and parallel residual adapters, regularization strategies, and parameter allocations, and empirically identify the ones that yield the highest compression. We show that, in order to maximize performance, it is necessary to adapt both shallow and deep layers of a deep network, but the required changes are very small. We also show that these universal parametrization are very effective for transfer learning, where they outperform traditional fine-tuning techniques.

****

Deep Density Clustering of Unconstrained Faces

Wei-An Lin, Jun-Cheng Chen, Carlos D. Castillo, Rama Chellappa; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8128-8137

In this paper, we consider the problem of grouping a collection of unconstrained face images in which the number of subjects is not known. We propose an unsupervised clustering algorithm called Deep Density Clustering (DDC) which is based on measuring density affinities between local neighborhoods in the feature space. By learning the minimal covering sphere for each neighborhood, information about the underlying structure is encapsulated. The encapsulation is also capable of locating high-density region of the neighborhood, which aids in measuring the neighborhood similarity. We theoretically show that the encapsulation asymptotically converges to a Parzen window density estimator. Our experiments show that DDC is a superior candidate for clustering unconstrained faces when the number of subjects is unknown. Unlike conventional linkage and density-based methods that are sensitive to the selection operating points, DDC attains more consistent and improved performance. Furthermore, the density-aware property reduces the difficulty in finding appropriate operating points.

****

Geometric Multi-Model Fitting With a Convex Relaxation Algorithm

Paul Amayo, Pedro Piniés, Lina M. Paz, Paul Newman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8138-8146

We propose a novel method for fitting multiple geometric models to multi-structural data via convex relaxation. Unlike greedy methods - which maximise the number of inliers - our approach efficiently searches for a soft assignment of points to geometric models by minimising the energy of the overall assignment. The inherently parallel nature of our approach, as compared to the sequential approach found in state-of-the-art energy minimisation techniques, allows for the elegant treatment of a scaling factor that occurs as the number of features in the data increases. This results in an energy minimisation that, per iteration, is as much as two orders of magnitude faster on comparable architectures thus bringing real-time, robust performance to a wider set of geometric multi-model fitting problems. We demonstrate the versatility of our approach on two canonical problems in estimating structure from images: plane extraction from RGB-D images and homography estimation from pairs of images. Our approach seamlessly adapts to the different metrics brought forth in these distinct problems. In both cases, we report results on publicly available data-sets that in most instances outperform th e state-of-the-art while simultaneously presenting run-times that are as much as an order of magnitude faster.

****

Fast and Robust Estimation for Unit-Norm Constrained Linear Fitting Problems

Daiki Ikami, Toshihiko Yamasaki, Kiyoharu Aizawa; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8147-8155

M-estimator using iteratively reweighted least squares (IRLS) is one of the best-known methods for robust estimation. However, IRLS is ineffective for robust unit-norm constrained linear fitting (UCLF) problems, such as fundamental matrix estimation because of a poor initial solution. We overcome this problem by developing a novel objective function and its optimization, named iteratively reweighted eigenvalues minimization (IREM). IREM is guaranteed to decrease the objective function and achieves fast convergence and high robustness. In robust fundamental matrix estimation, IREM performs approximately 5-500 times faster than random sampling consensus (RANSAC) while preserving comparable or superior robustness.

****

Importance Weighted Adversarial Nets for Partial Domain Adaptation

Jing Zhang, Zewei Ding, Wanqing Li, Philip Ogunbona; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8156-8164

This paper proposes an importance weighted adversarial nets-based method for unsupervised domain adaptation, specific for partial domain adaptation where the target domain has less number of classes compared to the source domain. Previous domain adaptation methods generally assume the identical label spaces, such that reducing the distribution divergence leads to feasible knowledge transfer. However, such an assumption is no longer valid in a more realistic scenario that requires adaptation from a larger and more diverse source domain to a smaller target domain with less number of classes. This paper extends the adversarial nets-based domain adaptation and proposes a novel adversarial nets-based partial domain adaptation method to identify the source samples that are potentially from the outlier classes and, at the same time, reduce the shift of shared classes between domains.

****

Efficient Subpixel Refinement With Symbolic Linear Predictors

Vincent Lui, Jonathon Geeves, Winston Yii, Tom Drummond; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8165-8173

We present an efficient subpixel refinement method using a learning-based approach called Linear Predictors. Firstly, we present a novel technique, called Symbolic Linear Predictors, which makes the learning step efficient for subpixel refinement. This makes our approach feasible for online applications without compromising accuracy, while taking advantage of the run-time efficiency of learning based approaches. Secondly, we show how Linear Predictors can be used to predict the expected alignment error, allowing us to use only the best keypoints in resource constrained applications. We show the efficiency and accuracy of our method through extensive experiments.

****

Scale-Recurrent Network for Deep Image Deblurring

Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, Jiaya Jia; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8174-8182

In single image deblurring, the ``coarse-to-fine'' scheme, i.e. gradually restoring the sharp image on different resolutions in a pyramid, is very successful in both traditional optimization-based methods and recent neural-network-based approaches. In this paper, we investigate this strategy and propose a Scale-recurrent Network (SRN-DeblurNet) for this deblurring task. Compared with the many recent learning-based approaches, it has a simpler network structure, a smaller number of parameters and is easier to train. We evaluate our method on large-scale deblurring datasets with complex motion. Results show that our method can produce better quality results than state-of-the-arts, both quantitatively and qualitatively.

****

DeblurGAN: Blind Motion Deblurring Using Conditional Adversarial Networks

Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, Ji■í Matas; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8183-8192

We present DeblurGAN, an end-to-end learned method for motion deblurring. The learning is based on a conditional GAN and the content loss . DeblurGAN achieves state-of-the art performance   both in the structural similarity measure and visual appearance. The quality of the deblurring model is also evaluated in a novel way on a real-world problem -- object detection on (de-)blurred images.   The method is 5 times faster than the closest competitor -- DeepDeblur. We also introduce a novel method for generating synthetic motion blurred images from sharp ones, allowing realistic dataset augmentation.   The model, code and the dataset are available at https://github.com/KupynOrest/DeblurGAN

****

A2-RL: Aesthetics Aware Reinforcement Learning for Image Cropping

Debang Li, Huikai Wu, Junge Zhang, Kaiqi Huang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8193-8201

Image cropping aims at improving the aesthetic quality of images by adjusting their composition. Most weakly supervised cropping methods (without bounding box supervision) rely on the sliding window mechanism. The sliding window mechanism requires fixed aspect ratios and limits the cropping region with arbitrary size. Moreover, the sliding window method usually produces tens of thousands of windows on the input image which is very time-consuming. Motivated by these challenges, we firstly formulate the aesthetic image cropping as a sequential decision-making process and propose a weakly supervised Aesthetics Aware Reinforcement Learning (A2-RL) framework to address this problem. Particularly, the proposed method develops an aesthetics aware reward function which especially benefits image cropping. Similar to human's decision making, we use a comprehensive state representation including both the current observation and the historical experience. We train the agent using the actor-critic architecture in an end-to-end manner. The agent is evaluated on several popular unseen cropping datasets. Experiment results show that our method achieves the state-of-the-art performance with much fewer candidate windows and much less time compared with previous weakly supervised methods.

****

Single Image Dehazing via Conditional Generative Adversarial Network

Runde Li, Jinshan Pan, Zechao Li, Jinhui Tang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8202-8211

In this paper, we present an algorithm to directly restore a clear image from a hazy image. This problem is highly ill-posed and most existing algorithms often use hand-crafted features, e.g., dark channel, color disparity, maximum contrast, to estimate transmission maps and then atmospheric lights. In contrast, we solve this problem based on a conditional generative adversarial network (cGAN), where the clear image is estimated by an end-to-end trainable neural network. Different from the generative network in basic cGAN, we propose an encoder and decoder architecture so that it can generate better results. To generate realistic clear images, we further modify the basic cGAN formulation by introducing the VGG features and a L_1-regularized gradient prior. We also synthesize a hazy dataset including indoor and outdoor scenes to train and evaluate the proposed algorithm. Extensive experimental results demonstrate that the proposed method performs favorably against the state-of-the-art methods on both synthetic dataset and real world hazy images.

****

On the Duality Between Retinex and Image Dehazing

Adrian Galdran, Aitor Alvarez-Gila, Alessandro Bria, Javier Vazquez-Corral, Marcelo Bertalmío; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8212-8221

Image dehazing deals with the removal of undesired loss of visibility in outdoor images due to the presence of fog. Retinex is a color vision model mimicking th

e ability of the Human Visual System to robustly discount varying illuminations when observing a scene under different spectral lighting conditions. Retinex has been widely explored in the computer vision literature for image enhancement and other related tasks. While these two problems are apparently unrelated, the goal of this work is to show that they can be connected by a simple linear relationship. Specifically, most Retinex-based algorithms have the characteristic feature of always increasing image brightness, which turns them into ideal candidates for effective image dehazing by directly applying Retinex to a hazy image whose intensities have been inverted. In this paper, we give theoretical proof that Retinex on inverted intensities is a solution to the image dehazing problem. Comprehensive qualitative and quantitative results indicate that several classical and modern implementations of Retinex can be transformed into competing image dehazing algorithms performing on pair with more complex fog removal methods, and can overcome some of the main challenges associated with this problem.
****

Arbitrary Style Transfer With Deep Feature Reshuffle
Shuyang Gu, Congliang Chen, Jing Liao, Lu Yuan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8222-8231
This paper introduces a novel method by reshuffling deep features (i.e., permuting the spacial locations of a feature map) of the style image for arbitrary style transfer. We theoretically prove that our new style loss based on reshuffle connects both global and local style losses respectively used by most parametric and non-parametric neural style transfer methods. This simple idea can effectively address the challenging issues in existing style transfer methods. On one hand, it can avoid distortions in local style patterns, and allow semantic-level transfer, compared with neural parametric methods. On the other hand, it can preserve globally similar appearance to the style image, and avoid wash-out artifacts, compared with neural non-parametric methods. Based on the proposed loss, we also present a progressive feature-domain optimization approach. The experiments show that our method is widely applicable to various styles, and produces better quality than existing methods.
****

Nonlocal Low-Rank Tensor Factor Analysis for Image Restoration
Xinyuan Zhang, Xin Yuan, Lawrence Carin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8232-8241
Low-rank signal modeling has been widely leveraged to capture non-local correlation in image processing applications. We propose a new method that employs low-rank tensor factor analysis for tensors generated by grouped image patches. The low-rank tensors are fed into the alternative direction multiplier method (ADMM) to further improve image reconstruction. The motivating application is compressive sensing (CS), and a deep convolutional architecture is adopted to approximate the expensive matrix inversion in CS applications. An iterative algorithm based on this low-rank tensor factorization strategy, called NLR-TFA, is presented in detail. Experimental results on noiseless and noisy CS measurements demonstrate the superiority of the proposed approach, especially at low CS sampling rates.
****

Avatar-Net: Multi-Scale Zero-Shot Style Transfer by Feature Decoration
Lu Sheng, Ziyi Lin, Jing Shao, Xiaogang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8242-8250
Zero-shot artistic style transfer is an important image synthesis problem aiming at transferring arbitrary style into content images. However, the trade-off between the generalization and efficiency in existing methods impedes a high quality zero-shot style transfer in real-time. In this paper, we resolve this dilemma and propose an efficient yet effective Avatar-Net that enables visually plausible multi-scale transfer for arbitrary style. The key ingredient of our method is a style decorator that makes up the content features by semantically aligned style features from an arbitrary style image, which does not only holistically match their feature distributions but also preserve detailed style patterns in the decorated features. By embedding this module into an image reconstruction network that fuses multi- scale style abstractions, the Avatar-Net renders multi-scale

stylization for any style image in one feed-forward pass. We demonstrate the state-of-the-art effectiveness and efficiency of the proposed method in generating high-quality stylized images, with a series of successive applications include multiple style integration, video stylization and etc.

****

Missing Slice Recovery for Tensors Using a Low-Rank Model in Embedded Space

Tatsuya Yokota, Burak Erem, Seyhmus Guler, Simon K. Warfield, Hidekata Hontani; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8251-8259

Let us consider a case where all of the elements in some continuous slices are missing in tensor data. In this case, the nuclear-norm and total variation regularization methods usually fail to recover the missing elements. The key problem is capturing some delay/shift-invariant structure. In this study, we consider a low-rank model in an embedded space of a tensor. For this purpose, we extend a delay embedding for a time series to a ``multi-way delay-embedding transform'' for a tensor, which takes a given incomplete tensor as the input and outputs a higher-order incomplete Hankel tensor. The higher-order tensor is then recovered by Tucker-based low-rank tensor factorization. Finally, an estimated tensor can be obtained by using the inverse multi-way delay embedding transform of the recovered higher-order tensor. Our experiments showed that the proposed method successfully recovered missing slices for some color images and functional magnetic resonance images.

****

Deep Semantic Face Deblurring

Ziyi Shen, Wei-Sheng Lai, Tingfa Xu, Jan Kautz, Ming-Hsuan Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8260-8269

In this paper, we present an effective and efficient face deblurring algorithm by exploiting semantic cues via deep convolutional neural networks (CNNs). As face images are highly structured and share several key semantic components (e.g., eyes and mouths), the semantic information of a face provides a strong prior for restoration. As such, we propose to incorporate global semantic priors as input and impose local structure losses to regularize the output within a multi-scale deep CNN. We train the network with perceptual and adversarial losses to generate photo-realistic results and develop an incremental training strategy to handle random blur kernels in the wild. Quantitative and qualitative evaluations demonstrate that the proposed face deblurring algorithm restores sharp images with more facial details and performs favorably against state-of-the-art methods in terms of restoration quality, face recognition and execution speed.

****

GraphBit: Bitwise Interaction Mining via Deep Reinforcement Learning

Yueqi Duan, Ziwei Wang, Jiwen Lu, Xudong Lin, Jie Zhou; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8270-8279

In this paper, we propose a GraphBit method to learn deep binary descriptors in a directed acyclic graph unsupervisedly, representing bitwise interactions as edges between the nodes of bits. Conventional binary representation learning methods enforce each element to be binarized into zero or one. However, there are elements lying in the boundary which suffer from doubtful binarization as ``ambiguous bits''. Ambiguous bits fail to collect effective information for confident binarization, which are unreliable and sensitive to noise. We argue that there are implicit inner relationships between bits in binary descriptors, where the related bits can provide extra instruction as prior knowledge for ambiguity elimination. Specifically, we design a deep reinforcement learning model to learn the structure of the graph for bitwise interaction mining, reducing the uncertainty of binary codes by maximizing the mutual information with inputs and related bits, so that the ambiguous bits receive additional instruction from the graph for confident binarization. Due to the reliability of the proposed binary codes with bitwise interaction, we obtain an average improvement of 9.64%, 8.84% and 3.22% on the CIFAR-10, Brown and HPatches datasets respectively compared with the state

-of-the-art unsupervised binary descriptors.

****

Recurrent Saliency Transformation Network: Incorporating Multi-Stage Visual Cues for Small Organ Segmentation

Qihang Yu, Lingxi Xie, Yan Wang, Yuyin Zhou, Elliot K. Fishman, Alan L. Yuille;
Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8280-8289

We aim at segmenting small organs (e.g., the pancreas) from abdominal CT scans. As the target often occupies a relatively small region in the input image, deep neural networks can be easily confused by the complex and variable background. To alleviate this, researchers proposed a coarse-to-fine approach, which used prediction from the first (coarse) stage to indicate a smaller input region for the second (fine) stage. Despite its effectiveness, this algorithm dealt with two stages individually, which lacked optimizing a global energy function, and limited its ability to incorporate multi-stage visual cues. Missing contextual information led to unsatisfying convergence in iterations, and that the fine stage sometimes produced even lower segmentation accuracy than the coarse stage. This paper presents a Recurrent Saliency Transformation Network. The key innovation is a saliency transformation module, which repeatedly converts the segmentation probability map from the previous iteration as spatial weights and applies these weights to the current iteration. This brings us two-fold benefits. In training, it allows joint optimization over the deep networks dealing with different input scales. In testing, it propagates multi-stage visual information throughout iterations to improve segmentation accuracy. Experiments in the NIH pancreas segmentation dataset demonstrate the state-of-the-art accuracy, which outperforms the previous best by an average of over 2%. Much higher accuracies are also reported on several small organs in a larger dataset collected by ourselves. In addition, our approach enjoys better convergence properties, making it more efficient and reliable in practice.

****

Thoracic Disease Identification and Localization With Limited Supervision

Zhe Li, Chong Wang, Mei Han, Yuan Xue, Wei Wei, Li-Jia Li, Li Fei-Fei; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8290-8299

Accurate identification and localization of abnormalities from radiology images play an integral part in clinical diagnosis and treatment planning. Building a highly accurate prediction model for these tasks usually requires a large number of images manually annotated with labels and finding sites of abnormalities. In reality, however, such annotated data are expensive to acquire, especially the ones with location annotations. We need methods that can work well with only a small amount of location annotations. To address this challenge, we present a unified approach that simultaneously performs disease identification and localization through the same underlying model for all images.We demonstrate that our approach can effectively leverage both class information as well as limited location annotation, and significantly outperforms the comparative reference baseline in both classification and localization tasks.

****

Quantization of Fully Convolutional Networks for Accurate Biomedical Image Segmentation

Xiaowei Xu, Qing Lu, Lin Yang, Sharon Hu, Danny Chen, Yu Hu, Yiyu Shi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8300-8308

With pervasive applications of medical imaging in healthcare, biomedical image segmentation plays a central role in quantitative analysis, clinical diagnosis, and medical intervention. Since manual annotation suffers limited reproducibility, arduous efforts, and excessive time, automatic segmentation is desired to process increasingly larger scale histopathological data. Recently, deep neural networks (DNNs), particularly fully convolutional networks (FCNs), have been widely applied to biomedical image segmentation, attaining much improved performance. At the same time, quantization of DNNs has become an active research topic, which aims to represent weights with less memory (precision) to considerably reduce memory and computation requirements of DNNs while maintaining acceptable accuracy. In this paper, we apply quantization techniques to FCNs for accurate biomedical image segmentation. Unlike existing literature on quantization which primarily targets memory and computation complexity reduction, we apply quantization as a method to reduce overfitting in FCNs for better accuracy. Specifically, we focus on a state-of-the-art segmentation framework, suggestive annotation [22], which judiciously extracts representative annotation samples from the original training dataset, obtaining an effective small-sized balanced training dataset. We develop two new quantization processes for this framework: (1) suggestive annotation with quantization for highly representative training samples, and (2) network training with quantization for high accuracy. Extensive experiments on the MICCAI Gland dataset show that both quantization processes can improve the segmentation performance, and our proposed method exceeds the current state-of-the-art performance by up to 1%. In addition, our method have a reduction of up to 6.4x on memory usage.

****

Visual Feature Attribution Using Wasserstein GANs

Christian F. Baumgartner, Lisa M. Koch, Kerem Can Tezcan, Jia Xi Ang, Ender Konukoglu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8309-8319

Attributing the pixels of an input image to a certain category is an important and well-studied problem in computer vision, with applications ranging from weakly supervised localisation to understanding hidden effects in the data. In recent years, approaches based on interpreting a previously trained neural network classifier have become the de facto state-of-the-art and are commonly used on medical as well as natural image datasets. In this paper, we discuss a limitation of these approaches which may lead to only a subset of the category specific features being detected. To address this problem we develop a novel feature attribution technique based on Wasserstein Generative Adversarial Networks (WGAN), which does not suffer from this limitation. We show that our proposed method performs substantially better than the state-of-the-art for visual attribution on a synthetic dataset and on real 3D neuroimaging data from patients with mild cognitive impairment (MCI) and Alzheimer's disease (AD). For AD patients the method produces compellingly realistic disease effect maps which are very close to the observed effects.

****

Total Capture: A 3D Deformation Model for Tracking Faces, Hands, and Bodies

Hanbyul Joo, Tomas Simon, Yaser Sheikh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8320-8329

We present a unified deformation model for the markerless capture of multiple scales of human movement, including facial expressions, body motion, and hand gestures. An initial model is generated by locally stitching together models of the individual parts of the human body, which we refer to as the ``Frankenstein'' model. This model enables the full expression of part movements, including face and hands by a single seamless model. Using a large-scale capture of people wearing everyday clothes, we optimize the Frankenstein model to create ``Adam". Adam is a model that shares the same skeleton hierarchy as the initial model, but can express hair and clothing geometry, making it directly usable for fitting people as they normally appear in everyday life. Finally, we demonstrate the use of these models for total motion tracking method, simultaneously capturing the large-scale body movements and the subtle face and hand motion of a social group of people.

****

Augmented Skeleton Space Transfer for Depth-Based Hand Pose Estimation

Seungryul Baek, Kwang In Kim, Tae-Kyun Kim; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8330-8339

Crucial to the success of training a depth-based 3D hand pose estimator (HPE) is the availability of comprehensive datasets covering diverse camera perspectives, shapes, and pose variations. However, collecting such annotated datasets is ch allenging. We propose to complete existing databases by generating new database entries. The key idea is to synthesize data in the skeleton space (instead of doing so in the depth-map space) which enables an easy and intuitive way of manipulating data entries. Since the skeleton entries generated in this way do not have the corresponding depth map entries, we exploit them by training a separate hand pose generator (HPG) which synthesizes the depth map from the skeleton entries. By training the HPG and HPE in a single unified optimization framework enforcing that 1) the HPE agrees with the paired depth and skeleton entries; and 2) the HPG-HPE combination satisfies the cyclic consistency (both the input and the output of HPG-HPE are skeletons) observed via the newly generated unpaired skeletons, our algorithm constructs a HPE which is robust to variations that go beyond the coverage of the existing database. Our training algorithm adopts the generative adversarial networks (GAN) training process. As a by-product, we obtain a hand pose discriminator (HPD) that is capable of picking out realistic hand poses. Our algorithm exploits this capability to refine the initial skeleton estimates in testing, further improving the accuracy. We test our algorithm on four challenging benchmark datasets (ICVL, MSRA, NYU and Big Hand 2.2M datasets) and demonstrate that our approach outperforms or is on par with state-of-the-art methods quantitatively and qualitatively.

****

Synthesizing Images of Humans in Unseen Poses

Guha Balakrishnan, Amy Zhao, Adrian V. Dalca, Frédo Durand, John Guttag; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8340-8348

We address the computational problem of novel human pose synthesis. Given an image of a person and a desired pose, we produce a depiction of that person in that pose, retaining the appearance of both the person and background. We present a modular generative neural network that synthesizes unseen poses using training pairs of images and poses taken from human action videos. Our network separates a scene into different body part and background layers, moves body parts to new locations and refines their appearances, and composites the new foreground with a hole-filled background. These subtasks, implemented with separate modules, are trained jointly using only a single target image as a supervised label. We use an adversarial discriminator to force our network to synthesize realistic details conditioned on pose. We demonstrate image synthesis results on three action classes: golf, yoga/workouts and tennis, and show that our method produces accurate results within action classes as well as across action classes. Given a sequence of desired poses, we also produce coherent videos of actions.

****

SSNet: Scale Selection Network for Online 3D Action Prediction

Jun Liu, Amir Shahroudy, Gang Wang, Ling-Yu Duan, Alex C. Kot; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8349-8358

In action prediction (early action recognition), the goal is to predict the class label of an ongoing action using its observed part so far. In this paper, we focus on online action prediction in streaming 3D skeleton sequences. A dilated convolutional network is introduced to model the motion dynamics in temporal dimension via a sliding window over the time axis. As there are significant temporal scale variations of the observed part of the ongoing action at different progress levels, we propose a novel window scale selection scheme to make our network focus on the performed part of the ongoing action and try to suppress the noise from the previous actions at each time step. Furthermore, an activation sharing scheme is proposed to deal with the overlapping computations among the adjacent steps, which allows our model to run more efficiently. The extensive experiments on two challenging datasets show the effectiveness of the proposed action prediction framework.

****

Detecting and Recognizing Human-Object Interactions

Georgia Gkioxari, Ross Girshick, Piotr Dollár, Kaiming He; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8359-

8367

To understand the visual world, a machine must not only recognize individual object instances but also how they interact. Humans are often at the center of such interactions and detecting human-object interactions is an important practical and scientific problem. In this paper, we address the task of detecting (human, verb, object) triplets in challenging everyday photos. We propose a novel model that is driven by a human-centric approach. Our hypothesis is that the appearance of a person -- their pose, clothing, action -- is a powerful cue for localizing the objects they are interacting with. To exploit this cue, our model learns to predict an action-specific density over target object locations based on the appearance of a detected person. Our model also jointly learns to detect people and objects, and by fusing these predictions it efficiently infers interaction triplets in a clean, jointly trained end-to-end system we call InteractNet. We validate our approach on the recently introduced Verbs in COCO (V-COCO) and HICO-DET datasets, where we show quantitatively compelling results.

****

Unsupervised Learning and Segmentation of Complex Activities From Video

Fadime Sener, Angela Yao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8368-8376

This paper presents a new method for unsupervised segmentation of complex activities from video into multiple steps, or sub-activities, without any textual input. We propose an iterative discriminative-generative approach which alternates between discriminatively learning the appearance of sub-activities from the videos' visual features to sub-activity labels and generatively modelling the temporal structure of sub-activities using a Generalized Mallows Model. In addition, we introduce a model for background to account for frames unrelated to the actual activities. Our approach is validated on the challenging Breakfast Actions and Inria Instructional Videos datasets and outperforms both unsupervised and weakly-supervised state of the art.

****

Unsupervised Training for 3D Morphable Model Regression

Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, William T. Freeman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8377-8386

We present a method for training a regression network from image pixels to 3D morphable model coordinates using only unlabeled photographs. The training loss is based on features from a facial recognition network, computed on-the-fly by rendering the predicted faces with a differentiable renderer. To make training from features feasible and avoid network fooling effects, we introduce three objectives: a batch distribution loss that encourages the output distribution to match the distribution of the morphable model, a loopback loss that ensures the network can correctly reinterpret its own output, and a multi-view identity loss that compares the features of the predicted 3D face and the input photograph from multiple viewing angles. We train a regression network using these objectives, a set of unlabeled photographs, and the morphable model itself, and demonstrate state-of-the-art results.

****

Video Based Reconstruction of 3D People Models

Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, Gerard Pons-Moll; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8387-8397

This paper describes how to obtain accurate 3D body models and texture of arbitrary people from a single, monocular video in which a person is moving. Based on a parametric body model, we present a robust processing pipeline achieving 3D model fits with 5mm accuracy also for clothed people. Our main contribution is a method to nonrigidly deform the silhouette cones corresponding to the dynamic human silhouettes, resulting in a visual hull in a common reference frame that enables surface reconstruction. This enables efficient estimation of a consensus 3D shape, texture and implanted animation skeleton based on a large number of frames. We present evaluation results for a number of test subjects and analyze overa ll performance. Requiring only a smartphone or webcam, our method enables everyone to create their own fully animatable digital double, e.g., for social VR applications or virtual try-on for online fashion shopping.

****

Pose-Guided Photorealistic Face Rotation

Yibo Hu, Xiang Wu, Bing Yu, Ran He, Zhenan Sun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8398-8406

Face rotation provides an effective and cheap way for data augmentation and representation learning of face recognition. It is a challenging generative learning problem due to the large pose discrepancy between two face images. This work focuses on flexible face rotation of arbitrary head poses, including extreme profile views. We propose a novel Couple-Agent Pose-Guided Generative Adversarial Network (CAPG-GAN) to generate both neutral and profile head pose face images. The head pose information is encoded by facial landmark heatmaps. It not only forms a mask image to guide the generator in learning process but also provides a flexible controllable condition during inference. A couple-agent discriminator is introduced to reinforce on the realism of synthetic arbitrary view faces. Besides the generator and conditional adversarial loss, CAPG-GAN further employs identity preserving loss and total variation regularization to preserve identity information and refine local textures respectively. Quantitative and qualitative experimental results on the Multi-PIE and LFW databases consistently show the superiority of our face rotation method over the state-of-the-art.

****

Mesoscopic Facial Geometry Inference Using Deep Neural Networks

Loc Huynh, Weikai Chen, Shunsuke Saito, Jun Xing, Koki Nagano, Andrew Jones, Paul Debevec, Hao Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8407-8416

We present a learning-based approach for synthesizing facial geometry at medium and fine scales from diffusely-lit facial texture maps. When applied to an image sequence, the synthesized detail is temporally coherent. Unlike current state-of-the-art methods, which assume "dark is deep", our model is trained with measured facial detail collected using polarized gradient illumination in a Light Stage. This enables us to produce plausible facial detail across the entire face, including where previous approaches may incorrectly interpret dark features as concavities such as at moles, hair stubble, and occluded pores. Instead of directly inferring 3D geometry, we propose to encode fine details in high-resolution displacement maps which are learned through a hybrid network adopting the state-of-the-art image-to-image translation network and super resolution network. To effectively capture geometric detail at both mid- and high frequencies, we factorize the learning into two separate sub-networks, enabling the full range of facial detail to be modeled. Results from our learning-based approach compare favorably with a high-quality active facial scanning technique, and require only a single passive lighting condition without a complex scanning setup.

****

Hand PointNet: 3D Hand Pose Estimation Using Point Sets

Liuhao Ge, Yujun Cai, Junwu Weng, Junsong Yuan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8417-8426

Convolutional Neural Network (CNN) has shown promising results for 3D hand pose estimation in depth images. Different from existing CNN-based hand pose estimation methods that take either 2D images or 3D volumes as the input, our proposed Hand PointNet directly processes the 3D point cloud that models the visible surface of the hand for pose regression. Taking the normalized point cloud as the input, our proposed hand pose regression network is able to capture complex hand structures and accurately regress a low dimensional representation of the 3D hand pose. In order to further improve the accuracy of fingertips, we design a fingertip refinement network that directly takes the neighboring points of the estimated fingertip location as input to refine the fingertip location. Experiments on three challenging hand pose datasets show that our proposed method outperforms state-of-the-art methods.

****

Seeing Voices and Hearing Faces: Cross-Modal Biometric Matching

Arsha Nagrani, Samuel Albanie, Andrew Zisserman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8427-8436

We introduce a seemingly impossible task: given only an audio clip of someone speaking, decide which of two face images is the speaker. In this paper we study this, and a number of related cross-modal tasks, aimed at answering the question: how much can we infer from the voice about the face and vice versa? We study this task “in the wild”, employing the datasets that are now publicly available for face recognition from static images (VGGFace) and speaker identification from audio (VoxCeleb). These provide training and testing scenarios for both static and dynamic testing of cross-modal matching. We make the following contributions: (i) we introduce CNN architectures for both binary and multi-way cross-modal face and audio matching; (ii) we compare dynamic testing (where video information is available, but the audio is not from the same video) with static testing (where only a single still image is available); and (iii) we use hu- man testing as a baseline to calibrate the difficulty of the task. We show that a CNN can indeed be trained to solve this task in both the static and dynamic scenarios, and is even well above chance on 10-way classification of the face given the voice. The CNN matches human performance on easy examples (e.g. different gender across faces) but exceeds human performance on more challenging examples (e.g. faces with the same gender, age and nationality).

****

Learning Monocular 3D Human Pose Estimation From Multi-View Images

Helge Rhodin, Jörg Spörri, Isinsu Katircioglu, Victor Constantin, Frédéric Meyer, Erich Müller, Mathieu Salzmann, Pascal Fua; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8437-8446

Accurate 3D human pose estimation from single images is possible with sophisticated deep-net architectures that have been trained on very large datasets. However, this still leaves open the problem of capturing motions for which no such database exists. Manual annotation is tedious, slow, and error-prone. In this paper, we propose to replace most of the annotations by the use of multiple views, at training time only. Specifically, we train the system to predict the same pose in all views. Such a consistency constraint is necessary but not sufficient to predict accurate poses. We therefore complement it with a supervised loss aiming to predict the correct pose in a small set of labeled images, and with a regularization term that penalizes drift from initial predictions. Furthermore, we propose a method to estimate camera pose jointly with human pose, which lets us utilize multi-view footage where calibration is difficult, e.g., for pan-tilt or moving handheld cameras. We demonstrate the effectiveness of our approach on established benchmarks, as well as on a new Ski dataset with rotating cameras and expert ski motion, for which annotations are truly hard to obtain.

****

Separating Style and Content for Generalized Style Transfer

Yexun Zhang, Ya Zhang, Wenbin Cai; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8447-8455

Neural style transfer has drawn broad attention in recent years. However, most existing methods aim to explicitly model the transformation between different styles, and the learned model is thus not generalizable to new styles. We here attempt to separate the representations for styles and contents, and propose a generalized style transfer network consisting of style encoder, content encoder, mixer and decoder. The style encoder and content encoder are used to extract the style and content factors from the style reference images and content reference images, respectively. The mixer employs a bilinear model to integrate the above two factors and finally feeds it into a decoder to generate images with target style and content. To separate the style features and content features, we leverage the conditional dependence of styles and contents given an image. During training, the encoder network learns to extract styles and contents from two sets of reference images in limited size, one with shared style and the other with shared content. This learning framework allows simultaneous style transfer among multiple styles and can be deemed as a special `multi-task' learning scenario. The enc oders are expected to capture the underlying features for different styles and contents which is generalizable to new styles and contents. For validation, we applied the proposed algorithm to the Chinese Typeface transfer problem. Extensive experiment results on character generation have demonstrated the effectiveness and robustness of our method.

****

TextureGAN: Controlling Deep Image Synthesis With Texture Patches

Wenqi Xian, Patsorn Sangkloy, Varun Agrawal, Amit Raj, Jingwan Lu, Chen Fang, Fisher Yu, James Hays; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8456-8465

In this paper, we investigate deep image synthesis guided by sketch, color, and texture. Previous image synthesis methods can be controlled by sketch and color strokes but we are the first to examine texture control. We allow a user to place a texture patch on a sketch at arbitrary locations and scales to control the desired output texture. Our generative network learns to synthesize objects consistent with these texture suggestions. To achieve this, we develop a local texture loss in addition to adversarial and content loss to train the generative network. We conduct experiments using sketches generated from real images and textures sampled from a separate texture database and results show that our proposed algorithm is able to generate plausible images that are faithful to user controls. Ablation studies show that our proposed pipeline can generate more realistic images than adapting existing methods directly.

****

Connecting Pixels to Privacy and Utility: Automatic Redaction of Private Information in Images

Tribhuvanesh Orekondy, Mario Fritz, Bernt Schiele; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8466-8475

Images convey a broad spectrum of personal information. If such images are shared on social media platforms, this personal information is leaked which conflicts with the privacy of depicted persons. Therefore, we aim for automated approaches to redact such private information and thereby protect privacy of the individual. By conducting a user study we find that obfuscating the image regions related to the private information leads to privacy while retaining utility of the images. Moreover, by varying the size of the regions different privacy-utility trade-offs can be achieved. Our findings argue for a "redaction by segmentation" paradigm.   Hence, we propose the first sizable dataset of private images "in the wild" annotated with pixel and instance level labels across a broad range of privacy classes. We present the first model for automatic redaction of diverse private information. It is effective at achieving various privacy-utility trade-offs within 83% of the performance of redactions based on ground-truth annotation.

****

MapNet: An Allocentric Spatial Memory for Mapping Environments

João F. Henriques, Andrea Vedaldi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8476-8484

Autonomous agents need to reason about the world beyond their instantaneous sensory input. Integrating information over time, however, requires switching from an egocentric representation of a scene to an allocentric one, expressed in the world reference frame. It must also be possible to update the representation dynamically, which requires localizing and registering the sensor with respect to it. In this paper, we develop a differentiable module that satisfies such requirements, while being robust, efficient, and suitable for integration in end-to-end deep networks. The module contains an allocentric spatial memory that can be accessed associatively by feeding to it the current sensory input, resulting in localization, and then updated using an LSTM or similar mechanism. We formulate efficient localization and registration of sensory information as a dual pair of convolution/deconvolution operators in memory space. The map itself is a 2.5D representation of an environment storing information that a deep neural network module learns to distill from RGBD input. The result is a map that contains multi-task information, different from classical approaches to mapping such as structure

-from-motion. We present results using synthetic mazes, a dataset of hours of recorded gameplay of the classic game Doom, and the very recent Active Vision Dataset of real images captured from a robot.

****

Accurate and Diverse Sampling of Sequences Based on a “Best of Many” Sample Objective

Apratim Bhattacharyya, Bernt Schiele, Mario Fritz; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8485-8493

For autonomous agents to successfully operate in the real world, anticipation of future events and states of their environment is a key competence. This problem has been formalized as a sequence extrapolation problem, where a number of observations are used to predict the sequence into the future. Real-world scenarios demand a model of uncertainty of such predictions, as predictions become increasingly uncertain -- in particular on long time horizons. While impressive results have been shown on point estimates, scenarios that induce multi-modal distributions over future sequences remain challenging. Our work addresses these challenges in a Gaussian Latent Variable model for sequence prediction. Our core contribution is a ``Best of Many'' sample objective that leads to more accurate and more diverse predictions that better capture the true variations in real-world sequence data. Beyond our analysis of improved model fit, our models also empirically outperform prior work on three diverse tasks ranging from traffic scenes to weather data.

****

VirtualHome: Simulating Household Activities via Programs

Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, Antonio Torralba; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8494-8502

In this paper, we are interested in modeling complex activities that occur in a typical household. We propose to use programs, i.e., sequences of atomic actions and interactions, as a high level representation of complex tasks. Programs are interesting because they provide a non-ambiguous representation of a task, and allow agents to execute them. However, nowadays, there is no database providing this type of information. Towards this goal, we first crowd-source programs for a variety of activities that happen in people's homes, via a game-like interface used for teaching kids how to code. Using the collected dataset, we show how we can learn to extract programs directly from natural language descriptions or from videos. We then implement the most common atomic (inter)actions in the Unity 3D game engine, and use our programs to "drive'' an artificial agent to execute tasks in a simulated household environment. Our VirtualHome simulator allows us to create a large activity video dataset with rich ground-truth, enabling training and testing of video understanding models. We further showcase examples of our agent performing tasks in our VirtualHome based on language

****

Generate to Adapt: Aligning Domains Using Generative Adversarial Networks

Swami Sankaranarayanan, Yogesh Balaji, Carlos D. Castillo, Rama Chellappa; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8503-8512

Domain Adaptation is an actively researched problem in Computer Vision. In this work, we propose an approach that leverages unsupervised data to bring the source and target distributions closer in a learned joint feature space. We accomplish this by inducing a symbiotic relationship between the learned embedding and a generative adversarial network. This is in contrast to methods which use the adversarial framework for realistic data generation and retraining deep models with such data. We demonstrate the strength and generality of our approach by performing experiments on three different tasks with varying levels of difficulty: (1) Digit classification (MNIST, SVHN and USPS datasets) (2) Object recognition using OFFICE dataset and (3) Domain adaptation from synthetic to real data. Our method achieves state-of-the art performance in most experimental settings and by far the only GAN-based method that has been shown to work well across different datasets such as OFFICE and DIGITS.

****

Multi-Agent Diverse Generative Adversarial Networks

Arnab Ghosh, Viveka Kulharia, Vinay P. Namboodiri, Philip H.S. Torr, Puneet K. Dokania; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8513-8521

We propose MAD-GAN, an intuitive generalization to the Generative Adversarial Networks (GANs) and its conditional variants to address the well known problem of mode collapse. First, MAD-GAN is a multi-agent GAN architecture incorporating multiple generators and one discriminator. Second, to enforce that different generators capture diverse high probability modes, the discriminator of MAD-GAN is designed such that along with finding the real and fake samples, it is also required to identify the generator that generated the given fake sample. Intuitively, to succeed in this task, the discriminator must learn to push different generators towards different identifiable modes. We perform extensive experiments on synthetic and real datasets and compare MAD-GAN with different variants of GAN. We show high quality diverse sample generations for challenging tasks such as image-to-image translation and face generation. In addition, we also show that MAD-GAN is able to disentangle different modalities when trained using highly challenging diverse-class dataset (e.g. dataset with images of forests, icebergs, and bedrooms). In the end, we show its efficacy on the unsupervised feature representation task.

****

A PID Controller Approach for Stochastic Optimization of Deep Networks

Wangpeng An, Haoqian Wang, Qingyun Sun, Jun Xu, Qionghai Dai, Lei Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8522-8531

Deep neural networks have demonstrated their power in many computer vision applications. State-of-the-art deep architectures such as VGG, ResNet, and DenseNet are mostly optimized by the SGD-Momentum algorithm, which updates the weights by considering their past and current gradients. Nonetheless, SGD-Momentum suffers from the overshoot problem, which hinders the convergence of network training. Inspired by the prominent success of proportional-integral-derivative (PID) controller in automatic control, we propose a PID approach for accelerating deep network optimization. We first reveal the intrinsic connections between SGD-Momentum and PID based controller, then present the optimization algorithm which exploits the past, current, and change of gradients to update the network parameters. The proposed PID method reduces much the overshoot phenomena of SGD-Momentum, and it achieves up to 50% acceleration on popular deep network architectures with competitive accuracy, as verified by our experiments on the benchmark datasets including CIFAR10, CIFAR100, and Tiny-ImageNet.

****

“Learning-Compression” Algorithms for Neural Net Pruning

Miguel Á. Carreira-Perpiñán, Yerlan Idelbayev; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8532-8541

Pruning a neural net consists of removing weights without degrading its performance. This is an old problem of renewed interest because of the need to compress ever larger nets so they can run in mobile devices. Pruning has been traditionally done by ranking or penalizing weights according to some criterion (such as magnitude), removing low-ranked weights and retraining the remaining ones. We formulate pruning as an optimization problem of finding the weights that minimize the loss while satisfying a pruning cost condition. We give a generic algorithm to solve this which alternates "learning" steps that optimize a regularized, data-dependent loss and "compression" steps that mark weights for pruning in a data-independent way. Magnitude thresholding arises naturally in the compression step, but unlike existing magnitude pruning approaches, our algorithm explores subsets of weights rather than committing irrevocably to a specific subset from the beginning. It is also able to learn automatically the best number of weights to prune in each layer of the net without incurring an exponentially costly model selection. Using a single pruning-level user parameter, we achieve state-of-the-art pruning in nets of various sizes.

****

Large-Scale Distance Metric Learning With Uncertainty

Qi Qian, Jiasheng Tang, Hao Li, Shenghuo Zhu, Rong Jin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8542-8550

Distance metric learning (DML) has been studied extensively in the past decades for its superior performance with distance-based algorithms. Most of the existing methods propose to learn a distance metric with pairwise or triplet constraints. However, the number of constraints is quadratic or even cubic in the number of the original examples, which makes it challenging for DML to handle the large-scale data set. Besides, the real-world data may contain various uncertainty, especially for the image data. The uncertainty can mislead the learning procedure and cause the performance degradation. By investigating the image data, we find that the original data can be observed from a small set of clean latent examples with different distortions. In this work, we propose the margin preserving metric learning framework to learn the distance metric and latent examples simultaneously. By leveraging the ideal properties of latent examples, the training efficiency can be improved significantly while the learned metric also becomes robust to the uncertainty in the original data. Furthermore, we can show that the metric is learned from latent examples only, but it can preserve the large margin property even for the original data. The empirical study on the benchmark image data sets demonstrates the efficacy and efficiency of the proposed method.

****

Guide Me: Interacting With Deep Networks

Christian Rupprecht, Iro Laina, Nassir Navab, Gregory D. Hager, Federico Tombari; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8551-8561

Interaction and collaboration between humans and intelligent machines has become increasingly important as machine learning methods move into real-world applications that involve end users. While much prior work lies at the intersection of natural language and vision, such as image captioning or image generation from text descriptions, less focus has been placed on the use of language to guide or improve the performance of a learned visual processing algorithm. In this paper, we explore methods to flexibly guide a trained convolutional neural network through user input to improve its performance during inference. We do so by inserting a layer that acts as a spatio-semantic guide into the network. This guide is trained to modify the network's activations, either directly via an energy minimization scheme or indirectly through a recurrent model that translates human language queries to interaction weights. Learning the verbal interaction is fully automatic and does not require manual text annotations. We evaluate the method on two datasets, showing that guiding a pre-trained network can improve performance, and provide extensive insights into the interaction between the guide and the CNN.

****

Art of Singular Vectors and Universal Adversarial Perturbations

Valentin Khrulkov, Ivan Oseledets; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8562-8570

Vulnerability of Deep Neural Networks (DNNs) to adversarial attacks has been attracting a lot of attention in recent studies. It has been shown that for many state of the art DNNs performing image classification there exist universal adversarial perturbations --- image-agnostic perturbations mere addition of which to natural images with high probability leads to their misclassification. In this work we propose a new algorithm for constructing such universal perturbations. Our approach is based on computing the so-called (p, q)-singular vectors of the Jacobian matrices of hidden layers of a network. Resulting perturbations present interesting visual patterns, and by using only 64 images we were able to construct universal perturbations with more than 60 % fooling rate on the dataset consisting of 50000 images. We also investigate a correlation between the maximal singular value of the Jacobian matrix and the fooling rate of the corresponding singular vector, and show that the constructed perturbations generalize across networ ks.

****

Deflecting Adversarial Attacks With Pixel Deflection
Aaditya Prakash, Nick Moran, Solomon Garber, Antonella DiLillo, James Storer; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8571-8580

CNNs are poised to become integral parts of many critical systems. Despite their robustness to natural variations, image pixel values can be manipulated, via small, carefully crafted, imperceptible perturbations, to cause a model to misclassify images. We present an algorithm to process an image so that classification accuracy is significantly preserved in the presence of such adversarial manipulations. Image classifiers tend to be robust to natural noise, and adversarial attacks tend to be agnostic to object location. These observations motivate our strategy, which leverages model robustness to defend against adversarial perturbations by forcing the image to match natural image statistics. Our algorithm locally corrupts the image by redistributing pixel values via a process we term pixel deflection. A subsequent wavelet-based denoising operation softens this corruption, as well as some of the adversarial changes. We demonstrate experimentally that the combination of these techniques enables the effective recovery of the true class, against a variety of robust attacks. Our results compare favorably with current state-of-the-art defenses, without requiring retraining or modifying the CNN.

****

MovieGraphs: Towards Understanding Human-Centric Situations From Videos
Paul Vicol, Makarand Tapaswi, Lluís Castrejón, Sanja Fidler; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8581-8590

There is growing interest in artificial intelligence to build socially intelligent robots. This requires machines to have the ability to "read" people's emotions, motivations, and other factors that affect behavior. Towards this goal, we introduce a novel dataset called MovieGraphs which provides detailed, graph-based annotations of social situations depicted in movie clips. Each graph consists of several types of nodes, to capture who is present in the clip, their emotional and physical attributes, their relationships (i.e., parent/child), and the interactions between them. Most interactions are associated with topics that provide additional details, and reasons that give motivations for actions. In addition, most interactions and many attributes are grounded in the video with time stamps. We provide a thorough analysis of our dataset, showing interesting common-sense correlations between different social aspects of scenes, as well as across scenes over time. We propose a method for querying videos and text with graphs, and show that: 1) our graphs contain rich and sufficient information to summarize and localize each scene; and 2) subgraphs allow us to describe situations at an abstract level and retrieve multiple semantically relevant situations. We also propose methods for interaction understanding via ordering, and reason understanding. MovieGraphs is the first benchmark to focus on inferred properties of human-centric situations, and opens up an exciting avenue towards socially-intelligent AI agents.

****

SemStyle: Learning to Generate Stylised Image Captions Using Unaligned Text
Alexander Mathews, Lexing Xie, Xuming He; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8591-8600

Linguistic style is an essential part of written communication, with the power to affect both clarity and attractiveness. With recent advances in vision and language, we can start to tackle the problem of generating image captions that are both visually grounded and appropriately styled. Existing approaches either require styled training captions aligned to images or generate captions with low relevance. We develop a model that learns to generate visually relevant styled captions from a large corpus of styled text without aligned images. The core idea of this model, called SemStyle, is to separate semantics and style. One key component is a novel and concise semantic term representation generated using natural language processing techniques and frame semantics. In addition, we develop a unified language model that decodes sentences with diverse word choices and syntax for different styles. Evaluations, both automatic and manual, show captions from SemStyle preserve image semantics, are descriptive, and are style shifted. More broadly, this work provides possibilities to learn richer image descriptions from the plethora of linguistic data available on the web.

****

Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions

Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, Fredrik Kahl, Tomas Pajdla; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8601-8610

Visual localization enables autonomous vehicles to navigate in their surroundings and augmented reality applications to link virtual to real worlds. Practical visual localization approaches need to be robust to a wide variety of viewing condition, including day-night changes, as well as weather and seasonal variations, while providing highly accurate 6 degree-of-freedom (6DOF) camera pose estimates. In this paper, we introduce the first benchmark datasets specifically designed for analyzing the impact of such factors on visual localization. Using carefully created ground truth poses for query images taken under a wide variety of conditions, we evaluate the impact of various factors on 6DOF camera pose estimation accuracy through extensive experiments with state-of-the-art localization approaches. Based on our results, we draw conclusions about the difficulty of different conditions, showing that long-term localization is far from solved, and propose promising avenues for future work, including sequence-based localization approaches and the need for better local features. Our benchmark is available at visuallocalization.net.

****

IVQA: Inverse Visual Question Answering

Feng Liu, Tao Xiang, Timothy M. Hospedales, Wankou Yang, Changyin Sun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8611-8619

We propose the inverse problem of Visual question answering (iVQA), and explore its suitability as a benchmark for visuo-linguistic understanding. The iVQA task is to generate a question that corresponds to a given image and answer pair. Since the answers are less informative than the questions, and the questions have less learnable bias, an iVQA model needs to better understand the image to be successful than a VQA model. We pose question generation as a multi-modal dynamic inference process and propose an iVQA model that can gradually adjust its focus of attention guided by both a partially generated question and the answer. For evaluation, apart from existing linguistic metrics, we propose a new ranking metric. This metric compares the ground truth question's rank among a list of distractors, which allows the drawbacks of different algorithms and sources of error to be studied. Experimental results show that our model can generate diverse, grammatically correct and content correlated questions that match the given answer.

****

Unsupervised Person Image Synthesis in Arbitrary Poses

Albert Pumarola, Antonio Agudo, Alberto Sanfeliu, Francesc Moreno-Noguer; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8620-8628

We present a novel approach for synthesizing photo-realistic images of people in arbitrary poses using generative adversarial learning. Given an input image of a person and a desired pose represented by a 2D skeleton, our model renders the image of the same person under the new pose, synthesizing novel views of the parts visible in the input image and hallucinating those that are not seen. This problem has recently been addressed in a supervised manner, i.e., during training the ground truth images under the new poses are given to the network. We go beyond these approaches by proposing a fully unsupervised strategy. We tackle this challenging scenario by splitting the problem into two principal subtasks. Firs t, we consider a pose conditioned bidirectional generator that maps back the initially rendered image to the original pose, hence being directly comparable to the input image without the need to resort to any training image. Second, we devise a novel loss function that incorporates content and style terms, and aims at producing images of high perceptual quality. Extensive experiments conducted on the DeepFashion dataset demonstrate that the images rendered by our model are very close in appearance to those obtained by fully supervised approaches.

****

Learning Descriptor Networks for 3D Shape Synthesis and Analysis

Jianwen Xie, Zilong Zheng, Ruiqi Gao, Wenguan Wang, Song-Chun Zhu, Ying Nian Wu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8629-8638

This paper proposes a 3D shape descriptor network, which is a deep convolutional energy-based model, for modeling volumetric shape patterns. The maximum likelihood training of the model follows an "analysis by synthesis" scheme and can be interpreted as a mode seeking and mode shifting process. The model can synthesize 3D shape patterns by sampling from the probability distribution via MCMC such as Langevin dynamics. The model can be used to train a 3D generator network via MCMC teaching. The conditional version of the 3D shape descriptor net can be used for 3D object recovery and 3D object super-resolution. Experiments demonstrate that the proposed model can generate realistic 3D shape patterns and can be useful for 3D shape analysis.

****

Neural Kinematic Networks for Unsupervised Motion Retargetting

Ruben Villegas, Jimei Yang, Duygu Ceylan, Honglak Lee; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8639-8648

We propose a recurrent neural network architecture with a Forward Kinematics layer and cycle consistency based adversarial training objective for unsupervised motion retargetting. Our network captures the high-level properties of an input motion by the forward kinematics layer, and adapts them to a target character with different skeleton bone lengths (e.g., shorter, longer arms etc.). Collecting paired motion training sequences from different characters is expensive. Instead, our network utilizes cycle consistency to learn to solve the Inverse Kinematics problem in an unsupervised manner. Our method works online, i.e., it adapts the motion sequence on-the-fly as new frames are received. In our experiments, we use the Mixamo animation data to test our method for a variety of motions and characters and achieve state-of-the-art results. We also demonstrate motion retargetting from monocular human videos to 3D characters using an off-the-shelf 3D pose estimator.

****

Group Consistent Similarity Learning via Deep CRF for Person Re-Identification

Dapeng Chen, Dan Xu, Hongsheng Li, Nicu Sebe, Xiaogang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8649-8658

Person re-identification benefits greatly from deep neural networks (DNN) to learn accurate similarity metrics and robust feature embeddings. However, most of the current methods impose only local constraints for similarity learning. In this paper, we incorporate constraints on large image groups by combining the CRF with deep neural networks. The proposed method aims to learn the ``local similarity" metrics for image pairs while taking into account the dependencies from all the images in a group, forming ``group similarities". Our method involves multiple images to model the relationships among the local and global similarities in a unified CRF during training, while combines multi-scale local similarities as the predicted similarity in testing. We adopt an approximate inference scheme for estimating the group similarity, enabling end-to-end training. Extensive experiments demonstrate the effectiveness of our model that combines DNN and CRF for learning robust multi-scale local similarities. The overall results outperform those by state-of-the-arts with considerable margins on three widely-used benchmarks.

****

Learning Compositional Visual Concepts With Mutual Consistency

Yunye Gong, Srikrishna Karanam, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, Peter C. Doerschuk; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8659-8668

Compositionality of semantic concepts in image synthesis and analysis is appealing as it can help in decomposing known and generatively recomposing unknown data. For instance, we may learn concepts of changing illumination, geometry or albedo of a scene, and try to recombine them to generate physically meaningful, but unseen data for training and testing. In practice however we often do not have samples from the joint concept space available: We may have data on illumination change in one data set and on geometric change in another one without complete overlap. We pose the following question: How can we learn two or more concepts jointly from different data sets with mutual consistency where we do not have samples from the full joint space? We present a novel answer in this paper based on cyclic consistency over multiple concepts, represented individually by generative adversarial networks (GANs). Our method, ConceptGAN, can be understood as a drop in for data augmentation to improve resilience for real world applications. Qualitative and quantitative evaluations demonstrate its efficacy in generating semantically meaningful images, as well as one shot face verification as an example application.

****

NestedNet: Learning Nested Sparse Structures in Deep Neural Networks

Eunwoo Kim, Chanho Ahn, Songhwai Oh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8669-8678

Recently, there have been increasing demands to construct compact deep architectures to remove unnecessary redundancy and to improve the inference speed. While many recent works focus on reducing the redundancy by eliminating unneeded weight parameters, it is not possible to apply a single deep network for multiple devices with different resources. When a new device or circumstantial condition requires a new deep architecture, it is necessary to construct and train a new network from scratch. In this work, we propose a novel deep learning framework, called a nested sparse network, which exploits an n-in-1-type nested structure in a neural network. A nested sparse network consists of multiple levels of networks with a different sparsity ratio associated with each level, and higher level networks share parameters with lower level networks to enable stable nested learning. The proposed framework realizes a resource-aware versatile architecture as the same network can meet diverse resource requirements, i.e., anytime property. Moreover, the proposed nested network can learn different forms of knowledge in its internal networks at different levels, enabling multiple tasks using a single network, such as coarse-to-fine hierarchical classification. In order to train the proposed nested network, we propose efficient weight connection learning and channel and layer scheduling strategies. We evaluate our network in multiple tasks, including adaptive deep compression, knowledge distillation, and learning class hierarchy, and demonstrate that nested sparse networks perform competitively, but more efficiently, compared to existing methods.

****

Context Embedding Networks

Kun Ho Kim, Oisin Mac Aodha, Pietro Perona; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8679-8687

Low dimensional embeddings that capture the main variations of interest in collections of data are important for many applications. One way to construct these embeddings is to acquire estimates of similarity from the crowd. Similarity is a multi-dimensional concept that varies from individual to individual. However, existing models for learning crowd embeddings typically make simplifying assumptions such as all individuals estimate similarity using the same criteria, the list of criteria is known in advance, or that the crowd workers are not influenced by the data that they see. To overcome these limitations we introduce Context Embedding Networks (CENs). In addition to learning interpretable embeddings from images, CENs also model worker biases for different attributes along with the visual context i.e. the attributes highlighted by a set of images. Experiments on t hree noisy crowd annotated datasets show that modeling both worker bias and visual context results in more interpretable embeddings compared to existing approaches.

****

Iterative Learning With Open-Set Noisy Labels

Yisen Wang, Weiyang Liu, Xingjun Ma, James Bailey, Hongyuan Zha, Le Song, Shu-Tao Xia; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8688-8696

Large-scale datasets possessing clean label annotations are crucial for training Convolutional Neural Networks (CNNs). However, labeling large-scale data can be very costly and error-prone, and even high-quality datasets are likely to contain noisy (incorrect) labels. Existing works usually employ a closed-set assumption, whereby the samples associated with noisy labels possess a true class contained within the set of known classes in the training data. However, such an assumption is too restrictive for many applications, since samples associated with noisy labels might in fact possess a true class that is not present in the training data. We refer to this more complex scenario as the open-set noisy label problem and show that it is nontrivial in order to make accurate predictions. To address this problem, we propose a novel iterative learning framework for training CNNs on datasets with open-set noisy labels. Our approach detects noisy labels and learns deep discriminative features in an iterative fashion. To benefit from the noisy label detection, we design a Siamese network to encourage clean labels and noisy labels to be dissimilar. A reweighting module is also applied to simultaneously emphasize the learning from clean labels and reduce the effect caused by noisy labels. Experiments on CIFAR-10, ImageNet and real-world noisy (web-search) datasets demonstrate that our proposed model can robustly train CNNs in the presence of a high proportion of open-set as well as closed-set noisy labels.

****

Learning Transferable Architectures for Scalable Image Recognition

Barret Zoph, Vijay Vasudevan, Jonathon Shlens, Quoc V. Le; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8697-8710

Developing neural network image classification models often requires significant architecture engineering. In this paper, we study a method to learn the model architectures directly on the dataset of interest. As this approach is expensive when the dataset is large, we propose to search for an architectural building block on a small dataset and then transfer the block to a larger dataset. The key contribution of this work is the design of a new search space (which we call the "NASNet search space"") which enables transferability. In our experiments, we search for the best convolutional layer (or "cell") on the CIFAR-10 dataset and then apply this cell to the ImageNet dataset by stacking together more copies of this cell, each with their own parameters to design a convolutional architecture, which we name a "NASNet architecture". We also introduce a new regularization technique called ScheduledDropPath that significantly improves generalization in the NASNet models. On CIFAR-10 itself, a NASNet found by our method achieves 2.4% error rate, which is state-of-the-art. Although the cell is not searched for directly on ImageNet, a NASNet constructed from the best cell achieves, among the published works, state-of-the-art accuracy of 82.7% top-1 and 96.2% top-5 on ImageNet. Our model is 1.2% better in top-1 accuracy than the best human-invented architectures while having 9 billion fewer FLOPS -- a reduction of 28% in computational demand from the previous state-of-the-art model. When evaluated at different levels of computational cost, accuracies of NASNets exceed those of the state-of-the-art human-designed models. For instance, a small version of NASNet also achieves 74% top-1 accuracy, which is 3.1% better than equivalently-sized, state-of-the-art models for mobile platforms. Finally, the image features learned from image classification are generically useful and can be transferred to other computer vision problems. On the task of object detection, the learned features by NASNet used with the Faster-RCNN framework surpass state-of-the-art by 4.0% achieving 43.1% mAP on the COCO dataset.

****

SBNet: Sparse Blocks Network for Fast Inference
Mengye Ren, Andrei Pokrovsky, Bin Yang, Raquel Urtasun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8711-8720
Conventional deep convolutional neural networks (CNNs) apply convolution operators uniformly in space across all feature maps for hundreds of layers - this incurs a high computational cost for real-time applications. For many problems such as object detection and semantic segmentation, we are able to obtain a low-cost computation mask, either from a priori problem knowledge, or from a low-resolution segmentation network. We show that such computation masks can be used to reduce computation in the high-resolution main network. Variants of sparse activation CNNs have previously been explored on small-scale tasks and showed no degradation in terms of object classification accuracy, but often measured gains in terms of theoretical FLOPs without realizing a practical speed-up when compared to highly optimized dense convolution implementations. In this work, we leverage the sparsity structure of computation masks and propose a novel tiling-based sparse convolution algorithm. We verified the effectiveness of our sparse CNN on LiDAR-based 3D object detection, and we report significant wall-clock speed-ups compared to dense convolution without noticeable loss of accuracy.
****
Language-Based Image Editing With Recurrent Attentive Models
Jianbo Chen, Yelong Shen, Jianfeng Gao, Jingjing Liu, Xiaodong Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8721-8729
We investigate the problem of Language-Based Image Editing (LBIE). Given a source image and a natural language description, we want to generate a target image by editing the source image based on the description. We propose a generic modeling framework for two sub-tasks of LBIE: language-based image segmentation and image colorization. The framework uses recurrent attentive models to fuse image and language features. Instead of using a fixed step size, we introduce for each region of the image a termination gate to dynamically determine after each inference step whether to continue extrapolating additional information from the textual description. The effectiveness of the framework is validated on three datasets. First, we introduce a synthetic dataset, called CoSaL, to evaluate the end-to-end performance of our LBIE system. Second, we show that the framework leads to state-of-the-art performance on image segmentation on the ReferIt dataset. Third, we present the first language-based colorization result on the Oxford-102 Flowers dataset.
****
Net2Vec: Quantifying and Explaining How Concepts Are Encoded by Filters in Deep Neural Networks
Ruth Fong, Andrea Vedaldi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8730-8738
In an effort to understand the meaning of the intermediate representations captured by deep networks, recent papers have tried to associate specific semantic concepts to individual neural network filter responses, where interesting correlations are often found, largely by focusing on extremal filter responses. In this paper, we show that this approach can favor easy-to-interpret cases that are not necessarily representative of the average behavior of a representation. A more realistic but harder-to-study hypothesis is that semantic representations are distributed, and thus filters must be studied in conjunction. In order to investigate this idea while enabling systematic visualization and quantification of multiple filter responses, we introduce the Net2Vec framework, in which semantic concepts are mapped to vectorial embeddings based on corresponding filter responses. By studying such embeddings, we are able to show that 1., in most cases, multiple filters are required to code for a concept, that 2., often filters are not concept specific and help encode multiple concepts, and that 3., compared to single filter activations, filter embeddings are able to better characterize the meaning of a representation and its relationship to other concepts.
****

End-to-End Dense Video Captioning With Masked Transformer

Luowei Zhou, Yingbo Zhou, Jason J. Corso, Richard Socher, Caiming Xiong; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8739-8748

Dense video captioning aims to generate text descriptions for all events in an untrimmed video. This involves both detecting and describing events. Therefore, all previous methods on dense video captioning tackle this problem by building two models, i.e. an event proposal and a captioning model, for these two sub-problems. The models are either trained separately or in alternation. This prevents direct influence of the language description to the event proposal, which is important for generating accurate descriptions. To address this problem, we propose an end-to-end transformer model for dense video captioning. The encoder encodes the video into appropriate representations. The proposal decoder decodes from the encoding with different anchors to form video event proposals. The captioning decoder employs a masking network to restrict its attention to the proposal event over the encoding feature. This masking network converts the event proposal to a differentiable mask, which ensures the consistency between the proposal and captioning during training. In addition, our model employs a self-attention mechanism, which enables the use of efficient non-recurrent structure during encoding and leads to performance improvements. We demonstrate the effectiveness of this end-to-end model on ActivityNet Captions and YouCookII datasets, where we achieved 10.12 and 6.58 METEOR score, respectively.

****

A Neural Multi-Sequence Alignment TeCHnique (NeuMATCH)

Pelin Dogan, Boyang Li, Leonid Sigal, Markus Gross; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8749-8758

The alignment of heterogeneous sequential data (video to text) is an important and challenging problem. Standard techniques for this task, including Dynamic Time Warping (DTW) and Conditional Random Fields (CRFs), suffer from inherent drawbacks. Mainly, the Markov assumption implies that, given the immediate past, future alignment decisions are independent of further history. The separation between similarity computation and alignment decision also prevents end-to-end training. In this paper, we propose an end-to-end neural architecture where alignment actions are implemented as moving data between stacks of Long Short-term Memory (LSTM) blocks. This flexible architecture supports a large variety of alignment tasks, including one-to-one, one-to-many, skipping unmatched elements, and (with extensions) non-monotonic alignment. Extensive experiments on semi-synthetic and real datasets show that our algorithm outperforms state-of-the-art baselines.

****

Path Aggregation Network for Instance Segmentation

Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, Jiaya Jia; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8759-8768

The way that information propagates in neural networks is of great importance. In this paper, we propose Path Aggregation Network (PANet) aiming at boosting information flow in proposal-based instance segmentation framework. Specifically, we enhance the entire feature hierarchy with accurate localization signals in lower layers by bottom-up path augmentation, which shortens the information path between lower layers and topmost feature. We present adaptive feature pooling, which links feature grid and all feature levels to make useful information in each level propagate directly to following proposal subnetworks. A complementary branch capturing different views for each proposal is created to further improve mask prediction. These improvements are simple to implement, with subtle extra computational overhead. Yet they are useful and make our PANet reach the 1st place in the COCO 2017 Challenge Instance Segmentation task and the 2nd place in Object Detection task without large-batch training. PANet is also state-of-the-art on MVD and Cityscapes.

****

The INaturalist Species Classification and Detection Dataset

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, Serge Belongie; Proceedings of the IEEE Conference on

Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8769-8778

Existing image classification datasets used in computer vision tend to have a uniform distribution of images across object categories. In contrast, the natural world is heavily imbalanced, as some species are more abundant and easier to photograph than others. To encourage further progress in challenging real world conditions we present the iNaturalist species classification and detection dataset, consisting of 859,000 images from over 5,000 different species of plants and animals. It features visually similar species, captured in a wide variety of situations, from all over the world. Images were collected with different camera types, have varying image quality, feature a large class imbalance, and have been verified by multiple citizen scientists. We discuss the collection of the dataset and present extensive baseline experiments using state-of-the-art computer vision classification and detection models. Results show that current non-ensemble based methods achieve only 67% top one classification accuracy, illustrating the difficulty of the dataset. Specifically, we observe poor results for classes with small numbers of training examples suggesting more attention is needed in low-shot learning.

****

Multimodal Explanations: Justifying Decisions and Pointing to the Evidence

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, Marcus Rohrbach; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8779-8788

Deep models that are both effective and explainable are desirable in many settings; prior explainable models have been unimodal, offering either image-based visualization of attention weights or text-based generation of post-hoc justifications. We propose a multimodal approach to explanation, and argue that the two modalities provide complementary explanatory strengths. We collect two new datasets to define and evaluate this task, and propose a novel model which can provide joint textual rationale generation and attention visualization. Our datasets define visual and textual justifications of a classification decision for activity recognition tasks (ACT-X) and for visual question answering tasks (VQA-X). We quantitatively show that training with the textual explanations not only yields better textual justification models, but also better localizes the evidence that supports the decision. We also qualitatively show cases where visual explanation is more insightful than textual explanation, and vice versa, supporting our thesis that multimodal explanation models offer significant benefits over unimodal approaches.

****

StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, Jaegul Choo; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8789-8797

Recent studies have shown remarkable success in image-to-image translation for two domains. However, existing approaches have limited scalability and robustness in handling more than two domains, since different models should be built independently for every pair of image domains. To address this limitation, we propose StarGAN, a novel and scalable approach that can perform image-to-image translations for multiple domains using only a single model. Such a unified model architecture of StarGAN allows simultaneous training of multiple datasets with different domains within a single network. This leads to StarGAN's superior quality of translated images compared to existing models as well as the novel capability of flexibly translating an input image to any desired target domain. We empirically demonstrate the effectiveness of our approach on a facial attribute transfer and a facial expression synthesis tasks.

****

High-Resolution Image Synthesis and Semantic Manipulation With Conditional GANs

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, Bryan Catanzaro; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8798-8807

We present a new method for synthesizing high-resolution photo-realistic images from semantic label maps using conditional generative adversarial networks (conditional GANs). Conditional GANs have enabled a variety of applications, but the results are often limited to low-resolution and still far from realistic. In this work, we generate 2048x1024 visually appealing results with a novel adversarial loss, as well as new multi-scale generator and discriminator architectures. Furthermore, we extend our framework to interactive visual manipulation with two additional features. First, we incorporate object instance segmentation information, which enables object manipulations such as removing/adding objects and changing the object category. Second, we propose a method to generate diverse results given the same input, allowing users to edit the object appearance interactively. Human opinion studies demonstrate that our method significantly outperforms existing methods, advancing both the quality and the resolution of deep image synthesis and editing.

****

Semi-Parametric Image Synthesis

Xiaojuan Qi, Qifeng Chen, Jiaya Jia, Vladlen Koltun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8808-8816

We present a semi-parametric approach to photographic image synthesis from semantic layouts. The approach combines the complementary strengths of parametric and nonparametric techniques. The nonparametric component is a memory bank of image segments constructed from a training set of images. Given a novel semantic layout at test time, the memory bank is used to retrieve photographic references that are provided as source material to a deep network. The synthesis is performed by a deep network that draws on the provided photographic material. Experiments on multiple semantic segmentation datasets show that the presented approach yields considerably more realistic images than recent purely parametric techniques.

****

BlockDrop: Dynamic Inference Paths in Residual Networks

Zuxuan Wu, Tushar Nagarajan, Abhishek Kumar, Steven Rennie, Larry S. Davis, Kristen Grauman, Rogerio Feris; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8817-8826

Very deep convolutional neural networks offer excellent recognition results, yet their computational expense limits their impact for many real-world applications. We introduce BlockDrop, an approach that learns to dynamically choose which layers of a deep network to execute during inference so as to best reduce total computation without degrading prediction accuracy.   Exploiting the robustness of Residual Networks (ResNets) to layer dropping, our framework selects on-the-fly which residual blocks to evaluate for a given novel image. In particular, given a pretrained ResNet, we train a policy network in an associative reinforcement learning setting for the dual reward of utilizing a minimal number of blocks while preserving recognition accuracy. We conduct extensive experiments on CIFAR and ImageNet. The results provide strong quantitative and qualitative evidence that these learned policies not only accelerate inference but also encode meaningful visual information. Built upon a ResNet-101 model, our method achieves a speedup of 20% on average, going as high as 36% for some images, while maintaining the same 76.4% top-1 accuracy on ImageNet.

****

Interpretable Convolutional Neural Networks

Quanshi Zhang, Ying Nian Wu, Song-Chun Zhu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8827-8836

This paper proposes a method to modify a traditional convolutional neural network (CNN) into an interpretable CNN, in order to clarify knowledge representations in high conv-layers of the CNN. In an interpretable CNN, each filter in a high conv-layer represents a specific object part. Our interpretable CNNs use the same training data as ordinary CNNs without a need for any annotations of object parts or textures for supervision. The interpretable CNN automatically assigns each filter in a high conv-layer with an object part during the learning process. We can apply our method to different types of CNNs with various structures. The explicit knowledge representation in an interpretable CNN can help people underst and the logic inside a CNN, i.e., what patterns are memorized by the CNN for prediction. Experiments have shown that filters in an interpretable CNN are more semantically meaningful than those in a traditional CNN. The code is available at https://github.com/zqs1022/interpretableCNN.

****

Deep Cross-Media Knowledge Transfer

Xin Huang, Yuxin Peng; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8837-8846

Cross-media retrieval is a research hotspot in multimedia area, which aims to perform retrieval across different media types such as image and text. The performance of existing methods usually relies on labeled data for model training. However, cross-media data is very labor consuming to collect and label, so how to transfer valuable knowledge in existing data to new data is a key problem towards application. For achieving the goal, this paper proposes deep cross-media knowledge transfer (DCKT) approach, which transfers knowledge from a large-scale cross-media dataset to promote the model training on another small-scale cross-media dataset. The main contributions of DCKT are: (1) Two-level transfer architecture is proposed to jointly minimize the media-level and correlation-level domain discrepancies, which allows two important and complementary aspects of knowledge to be transferred: intra-media semantic and inter-media correlation knowledge. It can enrich the training information and boost the retrieval accuracy. (2) Progressive transfer mechanism is proposed to iteratively select training samples with ascending transfer difficulties, via the metric of cross-media domain consistency with adaptive feedback. It can drive the transfer process to gradually reduce vast cross-media domain discrepancy, so as to enhance the robustness of model training. For verifying the effectiveness of DCKT, we take the large-scale dataset XMediaNet as source domain, and 3 widely-used datasets as target domain for cross-media retrieval. Experimental results show that DCKT achieves promising improvement on retrieval accuracy.

****

Interleaved Structured Sparse Convolutional Neural Networks

Guotian Xie, Jingdong Wang, Ting Zhang, Jianhuang Lai, Richang Hong, Guo-Jun Qi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8847-8856

In this paper, we study the problem of designing efficient convolutional neural network architectures with the interest in eliminating the redundancy in convolution kernels. In addition to structured sparse kernels, low-rank kernels and the product of low-rank kernels,the product of structured sparse kernels, which is a framework for interpreting the recently-developed interleaved group convolutions (IGC) and its variants (e.g. , Xception), has been attracting increasing interests.   Motivated by the observation that the convolutions contained in a group convolution in IGC can be further decomposed in the same manner, we present a modularized building block, {IGC-V2:}interleaved structured sparse convolutions. It generalizes interleaved group convolutions, which is composed of two structured sparse kernels, to the product of more structured sparse kernels, further eliminating the redundancy. We present the complementary condition and the balance condition to guide the design of structured sparse kernels, obtaining a balance between three aspects: model size and computation complexity and classification accuracy. Experimental results demonstrate the advantage on the balance between these three aspects compared to interleaved group convolutions and Xception and competitive performance with other state-of-the-art architecture design methods.

****

A Variational U-Net for Conditional Appearance and Shape Generation

Patrick Esser, Ekaterina Sutter, Björn Ommer; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8857-8866

Deep generative models have demonstrated great performance in image synthesis. However, results deteriorate in case of spatial deformations, since they generate images of objects directly, rather than modeling the intricate interplay of their inherent shape and appearance. We present a conditional U-Net for shape-guided image generation, conditioned on the output of a variational autoencoder for a ppearance. The approach is trained end-to-end on images, without requiring samples of the same object with varying pose or appearance. Experiments show that the model enables conditional image generation and transfer. Therefore, either shape or appearance can be retained from a query image, while freely altering the other. Moreover, appearance can be sampled due to its stochastic latent representation, while preserving shape. In quantitative and qualitative experiments on COCO, DeepFashion, shoes, Market-1501 and handbags, the approach demonstrates significant improvements over the state-of-the-art.

****

Detach and Adapt: Learning Cross-Domain Disentangled Deep Representation

Yen-Cheng Liu, Yu-Ying Yeh, Tzu-Chien Fu, Sheng-De Wang, Wei-Chen Chiu, Yu-Chiang Frank Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8867-8876

While representation learning aims to derive interpretable features for describing visual data, representation disentanglement further results in such features so that particular image attributes can be identified and manipulated. However, one cannot easily address this task without observing ground truth annotation for the training data. To address this problem, we propose a novel deep learning model of Cross-Domain Representation Disentangler (CDRD). By observing fully annotated source-domain data and unlabeled target-domain data of interest, our model bridges the information across data domains and transfers the attribute information accordingly. Thus, cross-domain joint feature disentanglement and adaptation can be jointly performed. In the experiments, we provide qualitative results to verify our disentanglement capability. Moreover, we further confirm that our model can be applied for solving classification tasks of unsupervised domain adaptation, and performs favorably against state-of-the-art image disentanglement and translation methods.

****

Learning Deep Structured Active Contours End-to-End

Diego Marcos, Devis Tuia, Benjamin Kellenberger, Lisa Zhang, Min Bai, Renjie Liao, Raquel Urtasun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8877-8885

The world is covered with millions of buildings, and precisely knowing each instance's position and extents is vital to a multitude of applications. Recently, automated building footprint segmentation models have shown superior detection accuracy thanks to the usage of Convolutional Neural Networks (CNN). However, even the latest evolutions struggle to precisely delineating borders, which often leads to geometric distortions and inadvertent fusion of adjacent building instances. We propose to overcome this issue by exploiting the distinct geometric properties of buildings. To this end, we present Deep Structured Active Contours (DSAC), a novel framework that integrates priors and constraints into the segmentation process, such as continuous boundaries, smooth edges, and sharp corners. To do so, DSAC employs Active Contour Models (ACM), a family of constraint- and prior-based polygonal models. We learn ACM parameterizations per instance using a CNN, and show how to incorporate all components in a structured output model, making DSAC trainable end-to-end. We evaluate DSAC on three challenging building instance segmentation datasets, where it compares favorably against state-of-the-art. Code will be made available.

****

Deep Learning Under Privileged Information Using Heteroscedastic Dropout

John Lambert, Ozan Sener, Silvio Savarese; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8886-8895

Unlike machines, humans learn through rapid, abstract model-building. The role of a teacher is not simply to hammer home right or wrong answers, but rather to provide intuitive comments, comparisons, and explanations to a pupil. This is what the Learning Under Privileged Information (LUPI) paradigm endeavors to model by utilizing extra knowledge only available during training. We propose a new LUPI algorithm specifically designed for Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). We propose to use a heteroscedastic dropout (ie. dropout with a varying variance) and make the variance of the dropout a functi on of privileged information. Intuitively, this corresponds to using the privileged information to control the uncertainty of the model output. We perform experiments using CNNs and RNNs for the tasks of image classification and machine translation. Our method significantly increases the sample efficiency during learning, resulting in higher accuracy with a large margin when the number of training examples is limited. We also theoretically justify the gains in sample efficiency by providing a generalization error bound decreasing with O(1/n), where n is the number of training examples, in an oracle case.

****

Smooth Neighbors on Teacher Graphs for Semi-Supervised Learning

Yucen Luo, Jun Zhu, Mengxi Li, Yong Ren, Bo Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8896-8905

The recently proposed self-ensembling methods have achieved promising results in deep semi-supervised learning, which penalize inconsistent predictions of unlabeled data under different perturbations. However, they only consider adding perturbations to each single data point, while ignoring the connections between data samples. In this paper, we propose a novel method, called Smooth Neighbors on Teacher Graphs (SNTG). In SNTG, a graph is constructed based on the predictions of the teacher model, i.e., the implicit self-ensemble of models. Then the graph serves as a similarity measure with respect to which the representations of "similar" neighboring points are learned to be smooth on the low-dimensional manifold. We achieve state-of-the-art results on semi-supervised learning benchmarks. The error rates are 9.89%, 3.99% for CIFAR-10 with 4000 labels, SVHN with 500 labels, respectively. In particular, the improvements are significant when the labels are fewer. For the non-augmented MNIST with only 20 labels, the error rate is reduced from previous 4.81% to 1.36%. Our method also shows robustness to noisy labels.

****

Interpret Neural Networks by Identifying Critical Data Routing Paths

Yulong Wang, Hang Su, Bo Zhang, Xiaolin Hu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8906-8914

Interpretability of a deep neural network aims to explain the rationale behind its decisions and enable the users to understand the intelligent agents, which has become an important issue due to its importance in practical applications. To address this issue, we develop a Distillation Guided Routing method, which is a flexible framework to interpret a deep neural network by identifying critical data routing paths and analyzing the functional processing behavior of the corresponding layers. Specifically, we propose to discover the critical nodes on the data routing paths during network inferring prediction for individual input samples by learning associated control gates for each layer's output channel. The routing paths can, therefore, be represented based on the responses of concatenated control gates from all the layers, which reflect the network's semantic selectivity regarding to the input patterns and more detailed functional process across different layer levels. Based on the discoveries, we propose an adversarial sample detection algorithm by learning a classifier to discriminate whether the critical data routing paths are from real or adversarial samples. Experiments demonstrate that our algorithm can effectively achieve high defense rate with minor training overhead.

****

Deep Spatio-Temporal Random Fields for Efficient Video Segmentation

Siddhartha Chandra, Camille Couprie, Iasonas Kokkinos; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8915-8924

In this work we introduce a time- and memory-efficient method for structured prediction that couples neuron decisions across both space at time. We show that we are able to perform exact and efficient inference on a densely connected spatio-temporal graph by capitalizing on recent advances on deep Gaussian Conditional Random Fields (GCRFs). Our method, called VideoGCRF is (a) efficient, (b) has a unique global minimum, and (c) can be trained end-to-end alongside contemporary deep networks for video understanding. We experiment with multiple connectivity patterns in the temporal domain, and present empirical improvements over strong baselines on the tasks of both semantic and instance segmentation of videos. Our implementation is based on the Caffe2 framework and will be available at https://github.com/siddharthachandra/gcrf-v3.0.

****

Customized Image Narrative Generation via Interactive Visual Question Generation and Answering

Andrew Shin, Yoshitaka Ushiku, Tatsuya Harada; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8925-8933

Image description task has been invariably examined in a static manner with qualitative presumptions held to be universally applicable, regardless of the scope or target of the description. In practice, however, different viewers may pay attention to different aspects of the image, and yield different descriptions or interpretations under various contexts. Such diversity in perspectives is difficult to derive with conventional image description techniques. In this paper, we propose a customized image narrative generation task, in which the users are interactively engaged in the generation process by providing answers to the questions. We further attempt to learn the user's interest via repeating such interactive stages, and to automatically reflect the interest in descriptions for new images. Experimental results demonstrate that our model can generate a variety of descriptions from single image that cover a wider range of topics than conventional models, while being customizable to the target user of interaction.

****

PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume

Deqing Sun, Xiaodong Yang, Ming-Yu Liu, Jan Kautz; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8934-8943

We present a compact but effective CNN model for optical flow, called PWC-Net. PWC-Net has been designed according to simple and well-established principles: pyramidal processing, warping, and the use of a cost volume. Cast in a learnable feature pyramid, PWC-Net uses the current optical flow estimate to warp the CNN features of the second image. It then uses the warped features and features of the first image to construct a cost volume, which is processed by a CNN to estimate the optical flow. PWC-Net is 17 times smaller in size and easier to train than the recent FlowNet2 model. Moreover, it outperforms all published optical flow methods on the MPI Sintel final pass and KITTI 2015 benchmarks, running at about 35 fps on Sintel resolution (1024x436) images. Our models are available on our project website.

****

Revisiting Deep Intrinsic Image Decompositions

Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, David Wipf; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8944-8952

While invaluable for many computer vision applications, decomposing a natural image into intrinsic reflectance and shading layers represents a challenging, underdetermined inverse problem. As opposed to strict reliance on conventional optimization or filtering solutions with strong prior assumptions, deep learning based approaches have also been proposed to compute intrinsic image decompositions when granted access to sufficient labeled training data. The downside is that current data sources are quite limited, and broadly speaking fall into one of two categories: either dense fully-labeled images in synthetic/narrow settings, or weakly-labeled data from relatively diverse natural scenes. In contrast to many previous learning-based approaches, which are often tailored to the structure of a particular dataset (and may not work well on others), we adopt core network structures that universally reflect loose prior knowledge regarding the intrinsic image formation process and can be largely shared across datasets. We then apply flexibly supervised loss layers that are customized for each source of ground truth labels. The resulting deep architecture achieves state-of-the-art results on all of the major intrinsic image benchmarks, and runs considerably faster than most at test time.

****

Multi-Cell Detection and Classification Using a Generative Convolutional Model

Florence Yellin, Benjamin D. Haeffele, Sophie Roth, René Vidal; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8953-8961

Detecting, counting, and classifying various cell types in images of human blood is important in many biomedical applications. However, these tasks can be very difficult due to the wide range of biological variability and the resolution limitations of many imaging modalities. This paper proposes a new approach to detecting, counting and classifying white blood cell populations in holographic images, which capitalizes on the fact that the variability in a mixture of blood cells is constrained by physiology. The proposed approach is based on a probabilistic generative model that describes an image of a population of cells as the sum of atoms from a convolutional dictionary of cell templates. The class of each template is drawn from a prior distribution that captures statistical information about blood cell mixtures. The parameters of the prior distribution are learned from a database of complete blood count results obtained from patients, and the cell templates are learned from images of purified cells from a single cell class using an extension of convolutional dictionary learning. Cell detection, counting and classification is then done using an extension of convolutional sparse coding that accounts for class proportion priors. This method has been successfully used to detect, count and classify white blood cell populations in holographic images of lysed blood obtained from 20 normal blood donors and 12 abnormal clinical blood discard samples. The error from our method is under 6.8% for all class populations, compared to errors of over 28.6% for all other methods tested.

****

Learning Spatial-Aware Regressions for Visual Tracking

Chong Sun, Dong Wang, Huchuan Lu, Ming-Hsuan Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8962-8970

In this paper, we analyze the spatial information of deep features, and propose two complementary regressions for robust visual tracking. First, we propose a kernelized ridge regression model wherein the kernel value is defined as the weighted sum of similarity scores of all pairs of patches between two samples. We show that this model can be formulated as a neural network and thus can be efficiently solved. Second, we propose a fully convolutional neural network with spatially regularized kernels, through which the filter kernel corresponding to each output channel is forced to focus on a specific region of the target. Distance transform pooling is further exploited to determine the effectiveness of each output channel of the convolution layer. The outputs from the kernelized ridge regression model and the fully convolutional neural network are combined to obtain the ultimate response. Experimental results on two benchmark datasets validate the effectiveness of the proposed method.

****

High Performance Visual Tracking With Siamese Region Proposal Network

Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, Xiaolin Hu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8971-8980

Visual object tracking has been a fundamental topic in recent years and many deep learning based trackers have achieved state-of-the-art performance on multiple benchmarks. However, most of these trackers can hardly get top performance with real-time speed. In this paper, we propose the Siamese region proposal network (Siamese-RPN) which is end-to-end trained off-line with large-scale image pairs. Specifically, it consists of Siamese subnetwork for feature extraction and region proposal subnetwork including the classification branch and regression branch. In the inference phase, the proposed framework is formulated as a local one-shot detection task. We can pre-compute the template branch of the Siamese subnetwork and formulate the correlation layers as trivial convolution layers to perform online tracking. Benefit from the proposal refinement, traditional multi-scale test and online fine-tuning can be discarded. The Siamese-RPN runs at 160 FPS while achieving leading performance in VOT2015, VOT2016 and VOT2017 real-time challenges.

****

LiteFlowNet: A Lightweight Convolutional Neural Network for Optical Flow Estimat ion

Tak-Wai Hui, Xiaoou Tang, Chen Change Loy; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8981-8989

FlowNet2, the state-of-the-art convolutional neural network (CNN) for optical flow estimation, requires over 160M parameters to achieve accurate flow estimation. In this paper we present an alternative network that attains performance on par with FlowNet2 on the challenging Sintel final pass and KITTI benchmarks, while being 30 times smaller in the model size and 1.36 times faster in the running speed. This is made possible by drilling down to architectural details that might have been missed in the current frameworks: (1) We present a more effective flow inference approach at each pyramid level through a lightweight cascaded network. It not only improves flow estimation accuracy through early correction, but also permits seamless incorporation of descriptor matching in our network. (2) We present a novel flow regularization layer to ameliorate the issue of outliers and vague flow boundaries by using a feature-driven local convolution. (3) Our network owns an effective structure for pyramidal feature extraction and embraces feature warping rather than image warping as practiced in FlowNet2. Our code and trained models are available at https://github.com/twhui/LiteFlowNet.

****

VITAL: VIsual Tracking via Adversarial Learning

Yibing Song, Chao Ma, Xiaohe Wu, Lijun Gong, Linchao Bao, Wangmeng Zuo, Chunhua Shen, Rynson W.H. Lau, Ming-Hsuan Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8990-8999

The tracking-by-detection framework consists of two stages, i.e., drawing samples around the target object in the first stage and classifying each sample as the target object or as background in the second stage. The performance of existing tracking-by-detection trackers using deep classification networks is limited by two aspects. First, the positive samples in each frame are highly spatially overlapped, and they fail to capture rich appearance variations. Second, there exists severe class imbalance between positive and negative samples. This paper presents the VITAL algorithm to address these two problems via adversarial learning. To augment positive samples, we use a generative network to randomly generate masks, which are applied to input features to capture a variety of appearance changes. With the use of adversarial learning, our network identifies the mask that maintains the most robust features of the target objects over a long temporal span. In addition, to handle the issue of class imbalance, we propose a high-order cost sensitive loss to decrease the effect of easy negative samples to facilitate training the classification network. Extensive experiments on benchmark datasets demonstrate that the proposed tracker performs favorably against state-of-the-art approaches.

****

Super SloMo: High Quality Estimation of Multiple Intermediate Frames for Video Interpolation

Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, Jan Kautz; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9000-9008

Given two consecutive frames, video interpolation aims at generating intermediate frame(s) to form both spatially and temporally coherent video sequences. While most existing methods focus on single-frame interpolation, we propose an end-to-end convolutional neural network for variable-length multi-frame video interpolation, where the motion interpretation and occlusion reasoning are jointly modeled. We start by computing bi-directional optical flow between the input images using a U-Net architecture. These flows are then linearly combined at each time step to approximate the intermediate bi-directional optical flows. These approximate flows, however, only work well in locally smooth regions and produce artifacts around motion boundaries. To address this shortcoming, we employ another U-Net to refine the approximated flow and also predict soft visibility maps. Finally, the two input images are warped and linearly fused to form each intermediate frame. By applying the visibility maps to the warped images before fusion, we exclude the contribution of occluded pixels to the interpolated intermediate frame to avoid artifacts. Since none of our learned network parameters are time-dependent, our approach is able to produce as many intermediate frames as needed. To train our network, we use 1,132 240-fps video clips, containing 300K individual video frames. Experimental results on several datasets, predicting different numbers of interpolated frames, demonstrate that our approach performs consistently better than existing methods.

****

Real-World Repetition Estimation by Div, Grad and Curl

Tom F. H. Runia, Cees G. M. Snoek, Arnold W. M. Smeulders; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9009-9017

We consider the problem of estimating repetition in video, such as performing push-ups, cutting a melon or playing violin. Existing work shows good results under the assumption of static and stationary periodicity. As realistic video is rarely perfectly static and stationary, the often preferred Fourier-based measurements is inapt. Instead, we adopt the wavelet transform to better handle non-static and non-stationary video dynamics. From the flow field and its differentials, we derive three fundamental motion types and three motion continuities of intrinsic periodicity in 3D. On top of this, the 2D perception of 3D periodicity considers two extreme viewpoints. What follows are 18 fundamental cases of recurrent perception in 2D. In practice, to deal with the variety of repetitive appearance, our theory implies measuring time-varying flow and its differentials (gradient, divergence and curl) over segmented foreground motion. For experiments, we introduce the new QUVA Repetition dataset, reflecting reality by including non-static and non-stationary videos. On the task of counting repetitions in video, we obtain favorable results compared to a deep learning alternative.

****

Recurrent Pixel Embedding for Instance Grouping

Shu Kong, Charless C. Fowlkes; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9018-9028

We introduce a differentiable, end-to-end trainable framework for solving pixel-level grouping problems such as instance segmentation consisting of two novel components. First, we regress pixels into a hyper-spherical embedding space so that pixels from the same group have high cosine similarity while those from different groups have similarity below a specified margin. We analyze the choice of embedding dimension and margin, relating them to theoretical results on the problem of distributing points uniformly on the sphere. Second, to group instances, we utilize a variant of mean-shift clustering, implemented as a recurrent neural network parameterized by kernel bandwidth. This recurrent grouping module is differentiable, enjoys convergent dynamics and probabilistic interpretability. Backpropagating the group-weighted loss through this module allows learning to focus on correcting embedding errors that won't be resolved during subsequent clustering. Our framework, while conceptually simple and theoretically abundant, is also practically effective and computationally efficient. We demonstrate substantial improvements over state-of-the-art instance segmentation for object proposal generation, as well as demonstrating the benefits of grouping loss for classification tasks such as boundary detection and semantic segmentation.

****

Deep Unsupervised Saliency Detection: A Multiple Noisy Labeling Perspective

Jing Zhang, Tong Zhang, Yuchao Dai, Mehrtash Harandi, Richard Hartley; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9029-9038

The success of current deep saliency detection methods heavily depends on the availability of large-scale supervision in the form of per-pixel labeling. Such supervision, while labor-intensive and not always possible, tends to hinder the generalization ability of the learned models. By contrast, traditional handcrafted features based unsupervised saliency detection methods, even though have been surpassed by the deep supervised methods, are generally dataset-independent and could be applied in the wild. This raises a natural question that ``Is it possible to learn saliency maps without using labeled data while improving the generali

zation ability?''. To this end, we present a novel perspective to unsupervised saliency detection through learning from multiple noisy labeling generated by ``weak'' and ``noisy'' unsupervised handcrafted saliency methods. Our end-to-end deep learning framework for unsupervised saliency detection consists of a latent saliency prediction module and a noise modeling module that work collaboratively and are optimized jointly. Explicit noise modeling enables us to deal with noisy saliency maps in a probabilistic way. Extensive experimental results on various benchmarking datasets show that our model not only outperforms all the unsupervised saliency methods with a large margin but also achieves comparable performance with the recent state-of-the-art supervised deep saliency methods.
****

Learning Intrinsic Image Decomposition From Watching the World
Zhengqi Li, Noah Snavely; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9039-9048
Single-view intrinsic image decomposition is a highly ill-posed problem, making learning from large amounts of data an attractive approach. However, it is difficult to collect ground truth training data at scale for intrinsic images. In this paper, we explore a different approach to learning intrinsic images: observing image sequences over time depicting the same scene under changing illumination, and learning single-view decompositions that are consistent with these changes. This approach allows us to learn without ground truth decompositions, and instead to exploit information available from multiple images. Our trained model can then be applied at test time to single views. We describe a new learning framework based on this idea, including new loss functions that can be efficiently evaluated over entire sequences. While prior learning-based intrinsic image methods achieve good performance on specific benchmarks, we show that our approach generalizes well to several diverse datasets, including MIT intrinsic images, Intrinsic Images in the Wild and Shading Annotations in the Wild.
****

TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-Rays
Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Ronald M. Summers; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9049-9058
Chest X-rays are one of the most common radiological examinations in daily clinical routines. Reporting thorax diseases using chest X-rays is often an entry-level task for radiologist trainees. Yet, reading a chest X-ray image remains a challenging job for learning-oriented machine intelligence, due to (1) shortage of large-scale machine-learnable medical image datasets, and (2) lack of techniques that can mimic the high-level reasoning of human radiologists that requires years of knowledge accumulation and professional training. In this paper, we show the clinical free-text radiological reports can be utilized as a priori knowledge for tackling these two key problems. We propose a novel Text-Image Embedding network (TieNet) for extracting the distinctive image and text representations. Multi-level attention models are integrated into an end-to-end trainable CNN-RNN architecture for highlighting the meaningful text words and image regions. We first apply TieNet to classify the chest X-rays by using both image features and text embeddings extracted from associated reports. The proposed auto-annotation framework achieves high accuracy (over 0.9 on average in AUCs) in assigning disease labels for our hand-label evaluation dataset. Furthermore, we transform the TieNet into a chest X-ray reporting system. It simulates the reporting process and can output disease classification and a preliminary report together. The classification results are significantly improved (6% increase on average in AUCs) compared to the state-of-the-art baseline on an unseen and hand-labeled dataset (OpenI).
****

Generating Synthetic X-Ray Images of a Person From the Surface Geometry
Brian Teixeira, Vivek Singh, Terrence Chen, Kai Ma, Birgi Tamersoy, Yifan Wu, Elena Balashova, Dorin Comaniciu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9059-9067

We present a novel framework that learns to predict human anatomy from body surface. Specifically, our approach generates a synthetic X-ray image of a person only from the person's surface geometry. Furthermore, the synthetic X-ray image is parametrized and can be manipulated by adjusting a set of body markers which are also generated during the X-ray image prediction. With the proposed framework, multiple synthetic X-ray images can easily be generated by varying surface geometry. By perturbing the parameters, several additional synthetic X-ray images can be generated from the same surface geometry. As a result, our approach offers a potential to overcome the training data barrier in the medical domain. This capability is achieved by learning a pair of networks - one learns to generate the full image from the partial image and a set of parameters, and the other learns to estimate the parameters given the full image. During training, the two networks are trained iteratively such that they would converge to a solution where the predicted parameters and the full image are consistent with each other. In addition to medical data enrichment, our framework can also be used for image completion as well as anomaly detection.

****

Gibson Env: Real-World Perception for Embodied Agents

Fei Xia, Amir R. Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, Silvio Savarese; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9068-9079

Perception and being active (having a certain level of motion freedom) are closely tied. Learning active perception and sensorimotor control in the physical world is cumbersome as existing algorithms are too slow to efficiently learn in real-time and robots are fragile and costly. This has given rise to learning in simulation which consequently casts a question on transferring to real-world. In this paper, we investigate learning a real-world perception for active agents, propose Gibson virtual environment for this purpose, and showcase a set of learned complex locomotion abilities. The primary characteristics of the learning environments, which transfer into the trained agents, are I) being from the real-world and reflecting its semantic complexity, II) having a mechanism to ensure no need to further domain adaptation prior to deployment of results in real-world, III) embodiment of the agent and making it subject to constraints of space and physics.

****

Reinforcement Cutting-Agent Learning for Video Object Segmentation

Junwei Han, Le Yang, Dingwen Zhang, Xiaojun Chang, Xiaodan Liang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9080-9089

Video object segmentation is a fundamental yet challenging task in computer vision community. In this paper, we formulate this problem as a Markov Decision Process, where agents are learned to segment object regions under a deep reinforcement learning framework. Essentially, learning agents for segmentation is nontrivial as segmentation is a nearly continuous decision-making process, where the number of the involved agents (pixels or superpixels) and action steps from the seed (super)pixels to the whole object mask might be incredibly huge. To overcome this difficulty, this paper simplifies the learning of segmentation agents to the learning of a cutting-agent, which only has a limited number of action units and can converge in just a few action steps. The basic assumption is that object segmentation mainly relies on the interaction between object regions and their context. Thus, with an optimal object (box) region and context (box) region, we can obtain the desirable segmentation mask through further inference. Based on this assumption, we establish a novel reinforcement cutting-agent learning framework, where the cutting-agent consists of a cutting-policy network and a cutting-execution network. The former learns policies for deciding optimal object-context box pair, while the latter executing the cutting function based on the inferred object-context box pair. With the collaborative interaction between the two networks, our method can achieve the outperforming VOS performance on two public benchmarks, which demonstrates the rationality of our assumption as well as the effectiveness of the proposed learning framework.

****

Feature Space Transfer for Data Augmentation

Bo Liu, Xudong Wang, Mandar Dixit, Roland Kwitt, Nuno Vasconcelos; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9090-9098

The problem of data augmentation in feature space is considered. A new architecture, denoted the FeATure TransfEr Network (FATTEN), is proposed for the modeling of feature trajectories induced by variations of object pose. This architecture exploits a parametrization of the pose manifold in terms of pose and appearance. This leads to a deep encoder/decoder network architecture, where the encoder factors into an appearance and a pose predictor. Unlike previous attempts at trajectory transfer, FATTEN can be efficiently trained end-to-end, with no need to train separate feature transfer functions. This is realized by supplying the decoder with information about a target pose and the use of a multi-task loss that penalizes category- and pose-mismatches. In result, FATTEN discourages discontinuous or non-smooth trajectories that fail to capture the structure of the pose manifold, and generalizes well on object recognition tasks involving large pose variation. Experimental results on the artificial ModelNet database show that it can successfully learn to map source features to target features of a desired pose, while preserving class identity. Most notably, by using feature space transfer for data augmentation (w.r.t. pose and depth) on SUN-RGBD objects, we demonstrate considerable performance improvements on one/few-shot object recognition in a transfer learning setup, compared to current state-of-the-art methods.

****

Analytic Expressions for Probabilistic Moments of PL-DNN With Gaussian Input

Adel Bibi, Modar Alfadly, Bernard Ghanem; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9099-9107

The outstanding performance of deep neural networks (DNNs), for the visual recognition task in particular, has been demonstrated on several large-scale benchmarks. This performance has immensely strengthened the line of re- search that aims to understand and analyze the driving reasons behind the effectiveness of these networks. One important aspect of this analysis has recently gained much attention, namely the reaction of a DNN to noisy input. This has spawned research on developing adversarial input attacks as well as training strategies that make DNNs more robust against these attacks. To this end, we derive in this pa- per exact analytic expressions for the first and second moments (mean and variance) of a small piecewise linear (PL) network (Affine, ReLU, Affine) subject to general Gaussian input. We experimentally show that these expressions are tight under simple linearizations of deeper PL-DNNs, especially popular architectures in the literature (e.g. LeNet and AlexNet). Extensive experiments on image classification show that these expressions can be used to study the behaviour of the output mean of the logits for each class, the interclass confusion and the pixel-level spatial noise sensitivity of the network. Moreover, we show how these expressions can be used to systematically construct targeted and non-targeted adversarial attacks.

****

Detail-Preserving Pooling in Deep Networks

Faraz Saeedan, Nicolas Weber, Michael Goesele, Stefan Roth; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9108-9116

Most convolutional neural networks use some method for gradually downscaling the size of the hidden layers. This is commonly referred to as pooling, and is applied to reduce the number of parameters, improve invariance to certain distortions, and increase the receptive field size. Since pooling by nature is a lossy process, it is crucial that each such layer maintains the portion of the activations that is most important for the network's discriminability. Yet, simple maximization or averaging over blocks, max or average pooling, or plain downsampling in the form of strided convolutions are the standard. In this paper, we aim to leverage recent results on image downscaling for the purposes of deep learning. Inspired by the human visual system, which focuses on local spatial changes, we pro pose detail-preserving pooling (DPP), an adaptive pooling method that magnifies spatial changes and preserves important structural detail. Importantly, its parameters can be learned jointly with the rest of the network. We analyze some of its theoretical properties and show its empirical benefits on several datasets and networks, where DPP consistently outperforms previous pooling approaches.

****

Rethinking Feature Distribution for Loss Functions in Image Classification

Weitao Wan, Yuanyi Zhong, Tianpeng Li, Jiansheng Chen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9117-9126

We propose a large-margin Gaussian Mixture (L-GM) loss for deep neural networks in classification tasks. Different from the softmax cross-entropy loss, our proposal is established on the assumption that the deep features of the training set follow a Gaussian Mixture distribution. By involving a classification margin and a likelihood regularization, the L-GM loss facilitates both a high classification performance and an accurate modeling of the training feature distribution. As such, the L-GM loss is superior to the softmax loss and its major variants in the sense that besides classification, it can be readily used to distinguish abnormal inputs, such as the adversarial examples, based on their features' likelihood to the training feature distribution. Extensive experiments on various recognition benchmarks like MNIST, CIFAR, ImageNet and LFW, as well as on adversarial examples demonstrate the effectiveness of our proposal.

****

Shift: A Zero FLOP, Zero Parameter Alternative to Spatial Convolutions

Bichen Wu, Alvin Wan, Xiangyu Yue, Peter Jin, Sicheng Zhao, Noah Golmant, Amir Gholaminejad, Joseph Gonzalez, Kurt Keutzer; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9127-9135

Neural networks rely on convolutions to aggregate spatial information. However, spatial convolutions are expensive in terms of model size and computation, both of which grow quadratically with respect to kernel size. In this paper, we present a parameter-free, FLOP-free "shift" operation as an alternative to spatial convolutions. We fuse shifts and point-wise convolutions to construct end-to-end trainable shift-based modules, with a hyperparameter characterizing the tradeoff between accuracy and efficiency. To demonstrate the operation's efficacy, we replace ResNet's 3x3 convolutions with shift-based modules for improved CIFAR-10 and CIFAR-100 accuracy using 60% fewer parameters; we additionally demonstrate the operation's resilience to parameter reduction on ImageNet, outperforming ResNet family members despite having millions fewer parameters. We further design a family of neural networks called ShiftNet, which achieve strong performance on classification, face verification and style transfer while demanding many fewer parameters.

****

Sketch-a-Classifier: Sketch-Based Photo Classifier Generation

Conghui Hu, Da Li, Yi-Zhe Song, Tao Xiang, Timothy M. Hospedales; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9136-9144

Contemporary deep learning techniques have made image recognition a reasonably reliable technology. However training effective photo classifiers typically takes numerous examples which limits image recognition's scalability and applicability to scenarios where images may not be available. This has motivated investigation into zero-shot learning, which addresses the issue via knowledge transfer from other modalities such as text. In this paper we investigate an alternative approach of synthesizing image classifiers: almost directly from a user's imagination, via free-hand sketch. This approach doesn't require the category to be nameable or describable via attributes as per zero-shot learning. We achieve this via training a model regression network to map from free-hand sketch space to the space of photo classifiers. It turns out that this mapping can be learned in a category-agnostic way, allowing photo classifiers for new categories to be synthesized by user with no need for annotated training photos. We also demonstrate that this modality of classifier generation can also be used to enhance the granularity of an existing photo classifier, or as a complement to name-based zero-shot

learning.
****

Light Field Intrinsics With a Deep Encoder-Decoder Network
Anna Alperovich, Ole Johannsen, Michael Strecke, Bastian Goldluecke; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9145-9154

We present a fully convolutional autoencoder for light fields, which jointly encodes stacks of horizontal and vertical epipolar plane images through a deep network of residual layers. The complex structure of the light field is thus reduced to a comparatively low-dimensional representation, which can be decoded in a variety of ways. The different pathways of upconvolution we currently support are for disparity estimation and separation of the lightfield into diffuse and specular intrinsic components. The key idea is that we can jointly perform unsupervised training for the autoencoder path of the network, and supervised training for the other decoders. This way, we find features which are both tailored to the respective tasks and generalize well to datasets for which only example light fields are available. We provide an extensive evaluation on synthetic light field data, and show that the network yields good results on previously unseen real world data captured by a Lytro Illum camera and various gantries.
****

Learning Generative ConvNets via Multi-Grid Modeling and Sampling
Ruiqi Gao, Yang Lu, Junpei Zhou, Song-Chun Zhu, Ying Nian Wu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9155-9164

This paper proposes a multi-grid method for learning energy-based generative ConvNet models of images. For each grid, we learn an energy-based probabilistic model where the energy function is defined by a bottom-up convolutional neural network (ConvNet or CNN). Learning such a model requires generating synthesized examples from the model. Within each iteration of our learning algorithm, for each observed training image, we generate synthesized images at multiple grids by initializing the finite-step MCMC sampling from a minimal 1 x 1 version of the training image. The synthesized image at each subsequent grid is obtained by a finite-step MCMC initialized from the synthesized image generated at the previous coarser grid. After obtaining the synthesized examples, the parameters of the models at multiple grids are updated separately and simultaneously based on the differences between synthesized and observed examples. We show that this multi-grid method can learn realistic energy-based generative ConvNet models, and it outperforms the original contrastive divergence (CD) and persistent CD.
****

Manifold Learning in Quotient Spaces
Éloi Mehr, André Lieutier, Fernando Sanchez Bermudez, Vincent Guitteny, Nicolas Thome, Matthieu Cord; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9165-9174

When learning 3D shapes we are usually interested in their intrinsic geometry rather than in their orientation. To deal with the orientation variations the usual trick consists in augmenting the data to exhibit all possible variability, and thus let the model learn both the geometry as well as the rotations. In this paper we introduce a new autoencoder model for encoding and synthesis of 3D shapes. To get rid of undesirable input variability our model learns a manifold in a quotient space of the input space. Typically, we propose to quotient the space of 3D models by the action of rotations. Thus, our quotient autoencoder allows to directly learn in the space of interest, ignoring side information. This is reflected in better performances on reconstruction and interpolation tasks, as our experiments show that our model outperforms a vanilla autoencoder on the well-known Shapenet dataset. Moreover, our model learns a rotation-invariant representation, leading to interesting results in shapes co-alignment. Finally, we extend our quotient autoencoder to quotient by non-rigid transformations.
****

Learning Intelligent Dialogs for Bounding Box Annotation
Ksenia Konyushkova, Jasper Uijlings, Christoph H. Lampert, Vittorio Ferrari; Pro

ceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9175-9184

We introduce Intelligent Annotation Dialogs for bounding box annotation. We train an agent to automatically choose a sequence of actions for a human annotator to produce a bounding box in a minimal amount of time. Specifically, we consider two actions: box verification, where the annotator verifies a box generated by an object detector, and manual box drawing. We explore two kinds of agents, one based on predicting the probability that a box will be positively verified, and the other based on reinforcement learning. We demonstrate that (1) our agents are able to learn efficient annotation strategies in several scenarios, automatically adapting to the image difficulty, the desired quality of the boxes, and the detector strength; (2) in all scenarios the resulting annotation dialogs speed up annotation compared to manual box drawing alone and box verification alone, while also outperforming any fixed combination of verification and drawing in most scenarios; (3) in a realistic scenario where the detector is iteratively re-trained, our agents evolve a series of strategies that reflect the shifting trade-off between verification and drawing as the detector grows stronger.

****

Boosting Adversarial Attacks With Momentum

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, Jianguo Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9185-9193

Deep neural networks are vulnerable to adversarial examples, which poses security concerns on these algorithms due to the potentially severe consequences. Adversarial attacks serve as an important surrogate to evaluate the robustness of deep learning models before they are deployed. However, most of existing adversarial attacks can only fool a black-box model with a low success rate. To address this issue, we propose a broad class of momentum-based iterative algorithms to boost adversarial attacks. By integrating the momentum term into the iterative process for attacks, our methods can stabilize update directions and escape from poor local maxima during the iterations, resulting in more transferable adversarial examples. To further improve the success rates for black-box attacks, we apply momentum iterative algorithms to an ensemble of models, and show that the adversarially trained models with a strong defense ability are also vulnerable to our black-box attacks. We hope that the proposed methods will serve as a benchmark for evaluating the robustness of various deep models and defense methods. With this method, we won the first places in NIPS 2017 Non-targeted Adversarial Attack and Targeted Adversarial Attack competitions.

****

NISP: Pruning Networks Using Neuron Importance Score Propagation

Ruichi Yu, Ang Li, Chun-Fu Chen, Jui-Hsin Lai, Vlad I. Morariu, Xintong Han, Mingfei Gao, Ching-Yung Lin, Larry S. Davis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9194-9203

To reduce the significant redundancy in deep Convolutional Neural Networks (CNNs), most existing methods prune neurons by only considering the statistics of an individual layer or two consecutive layers (e.g., prune one layer to minimize the reconstruction error of the next layer), ignoring the effect of error propagation in deep networks. In contrast, we argue that for a pruned network to retain its predictive power, it is essential to prune neurons in the entire neuron network jointly based on a unified goal: minimizing the reconstruction error of important responses in the ``final response layer" (FRL), which is the second-to-last layer before classification. Specifically, we apply feature ranking techniques to measure the importance of each neuron in the FRL, formulate network pruning as a binary integer optimization problem, and derive a closed-form solution to it for pruning neurons in earlier layers. Based on our theoretical analysis, we propose the Neuron Importance Score Propagation (NISP) algorithm to propagate the importance scores of final responses to every neuron in the network. The CNN is pruned by removing neurons with least importance, and it is then fine-tuned to recover its predictive power. NISP is evaluated on several datasets with multiple CNN models and demonstrated to achieve significant acceleration and compressio

n with negligible accuracy loss.
****

PointGrid: A Deep Network for 3D Shape Understanding
Truc Le, Ye Duan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9204-9214

This paper presents a new deep learning architecture called PointGrid that is designed for 3D model recognition from unorganized point clouds. The new architecture embeds the input point cloud into a 3D grid by a simple, yet effective, sampling strategy and directly learns transformations and features from their raw coordinates. The proposed method is an integration of point and grid, a hybrid model, that leverages the simplicity of grid-based approaches such as VoxelNet while avoid its information loss. PointGrid learns better global information compared with PointNet and is much simpler than PointNet++, Kd-Net, Oct-Net and O-CNN, yet provides comparable recognition accuracy. With experiments on popular shape recognition benchmarks, PointGrid demonstrates competitive performance over existing deep learning methods on both classification and segmentation.
****

Tell Me Where to Look: Guided Attention Inference Network
Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, Yun Fu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9215-9223

Weakly supervised learning with only coarse labels can obtain visual explanations of deep neural network such as attention maps by back-propagating gradients. These attention maps are then available as priors for tasks such as object localization and semantic segmentation. In one common framework we address three shortcomings of previous approaches in modeling such attention maps: We (1) make attention maps an explicit and natural component of the end-to-end training for the first time, (2) provide self-guidance directly on these maps by exploring supervision from the network itself to improve them, and (3) seamlessly bridge the gap between using weak and extra supervision if available. Despite its simplicity, experiments on the semantic segmentation task demonstrate the effectiveness of our methods. We clearly surpass the state-of-the-art on PASCAL VOC 2012 test and val. sets. Besides, the proposed framework provides a way not only explaining the focus of the learner but also feeding back with direct guidance towards specific tasks. Under mild assumptions our method can also be understood as a plug-in to existing weakly supervised learners to improve their generalization performance.
****

3D Semantic Segmentation With Submanifold Sparse Convolutional Networks
Benjamin Graham, Martin Engelcke, Laurens van der Maaten; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9224-9232

Convolutional networks are the de-facto standard for analyzing spatio-temporal data such as images, videos, and 3D shapes. Whilst some of this data is naturally dense (e.g., photos), many other data sources are inherently sparse. Examples include 3D point clouds that were obtained using a LiDAR scanner or RGB-D camera. Standard ``dense'' implementations of convolutional networks are very inefficient when applied on such sparse data. We introduce new sparse convolutional operations that are designed to process spatially-sparse data more efficiently, and use them to develop spatially-sparse convolutional networks. We demonstrate the strong performance of the resulting models, called submanifold sparse convolutional networks (SSCNs), on two tasks involving semantic segmentation of 3D point clouds. In particular, our models outperform all prior state-of-the-art on the test set of a recent semantic segmentation competition.
****

TOM-Net: Learning Transparent Object Matting From a Single Image
Guanying Chen, Kai Han, Kwan-Yee K. Wong; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9233-9241

This paper addresses the problem of transparent object matting. Existing image matting approaches for transparent objects often require tedious capturing proced ures and long processing time, which limit their practical use. In this paper, we first formulate transparent object matting as a refractive flow estimation problem. We then propose a deep learning framework, called TOM-Net, for learning the refractive flow. Our framework comprises two parts, namely a multi-scale encoder-decoder network for producing a coarse prediction, and a residual network for refinement. At test time, TOM-Net takes a single image as input, and outputs a matte (consisting of an object mask, an attenuation mask and a refractive flow field) in a fast feed-forward pass. As no off-the-shelf dataset is available for transparent object matting, we create a large-scale synthetic dataset consisting of 178K images of transparent objects rendered in front of images sampled from the Microsoft COCO dataset. We also collect a real dataset consisting of 876 samples using 14 transparent objects and 60 background images. Promising experimental results have been achieved on both synthetic and real data, which clearly demonstrate the effectiveness of our approach.

****

Translating and Segmenting Multimodal Medical Volumes With Cycle- and Shape-Consistency Generative Adversarial Network

Zizhao Zhang, Lin Yang, Yefeng Zheng; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9242-9251

Synthesized medical images have several important applications, e.g., as an intermedium in cross-modality image registration and as supplementary training samples to boost the generalization capability of a classifier. Especially, synthesized CT data can provide X-ray attenuation map for radiation therapy planning. In this work, we propose a generic cross-modality synthesis approach with the following targets: 1) synthesizing realistic looking 3D images using unpaired training data, 2) ensuring consistent anatomical structures, which could changed by geometric distortion in cross-modality synthesis and 3) improving volume segmentation by using synthetic data for modalities with limited training samples. We show that these goals can be achieved with an end-to-end 3D convolutional neural network (CNN) composed of mutually-beneficial generators and segmentors for image synthesis and segmentation tasks. The generators are trained with an adversarial loss, a cycle-consistency loss, and also a shape-consistency loss, which is supervised by segmentors, to reduce the geometric distortion. From the segmentation view, the segmentors are boosted by synthetic data from generators in an online manner. Generators and segmentors prompt each other alternatively in an end-to-end training fashion. With extensive experiments on a dataset including a total of 4,496 CT and MRI cardiovascular volumes, we show both tasks are beneficial to each other and coupling these two tasks results in better performance than solving them exclusively.

****

An Unsupervised Learning Model for Deformable Medical Image Registration

Guha Balakrishnan, Amy Zhao, Mert R. Sabuncu, John Guttag, Adrian V. Dalca; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9252-9260

We present a fast learning-based algorithm for deformable, pairwise 3D medical image registration. Current registration methods optimize an objective function independently for each pair of images, which can be time-consuming for large data. We define registration as a parametric function, and optimize its parameters given a set of images from a collection of interest. Given a new pair of scans, we can quickly compute a registration field by directly evaluating the function using the learned parameters. We model this function using a CNN, and use a spatial transform layer to reconstruct one image from another while imposing smoothness constraints on the registration field. The proposed method does not require supervised information such as ground truth registration fields or anatomical landmarks. We demonstrate registration accuracy comparable to state-of-the-art 3D image registration, while operating orders of magnitude faster in practice. Our method promises to significantly speed up medical image analysis and processing pipelines, while facilitating novel directions in learning-based registration and its applications. Our code is available at https://github.com/balakg/voxelmorph.

****

Deep Lesion Graphs in the Wild: Relationship Learning and Organization of Significant Radiology Image Findings in a Diverse Large-Scale Lesion Database

Ke Yan, Xiaosong Wang, Le Lu, Ling Zhang, Adam P. Harrison, Mohammadhadi Bagheri, Ronald M. Summers; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9261-9270

Radiologists in their daily work routinely find and annotate significant abnormalities on a large number of radiology images. Such abnormalities, or lesions, have collected over years and stored in hospitals' picture archiving and communication systems. However, they are basically unsorted and lack semantic annotations like type and location. In this paper, we aim to organize and explore them by learning a deep feature representation for each lesion. A large-scale and comprehensive dataset, DeepLesion, is introduced for this task. DeepLesion contains bounding boxes and size measurements of over 32K lesions. To model their similarity relationship, we leverage multiple supervision information including types, self-supervised location coordinates, and sizes. They require little manual annotation effort but describe useful attributes of the lesions. Then, a triplet network is utilized to learn lesion embeddings with a sequential sampling strategy to depict their hierarchical similarity structure. Experiments show promising qualitative and quantitative results on lesion retrieval, clustering, and classification. The learned embeddings can be further employed to build a lesion graph for various clinically useful applications. An algorithm for intra-patient lesion matching is proposed and validated with experiments.

****

Learning Distributions of Shape Trajectories From Longitudinal Datasets: A Hierarchical Model on a Manifold of Diffeomorphisms

Alexandre Bône, Olivier Colliot, Stanley Durrleman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9271-9280

We propose a method to learn a distribution of shape trajectories from longitudinal data, i.e. the collection of individual objects repeatedly observed at multiple time-points. The method allows to compute an average spatiotemporal trajectory of shape changes at the group level, and the individual variations of this trajectory both in terms of geometry and time dynamics. First, we formulate a non-linear mixed-effects statistical model as the combination of a generic statistical model for manifold-valued longitudinal data, a deformation model defining shape trajectories via the action of a finite-dimensional set of diffeomorphisms with a manifold structure, and an efficient numerical scheme to compute parallel transport on this manifold. Second, we introduce a MCMC-SAEM algorithm with a specific approach to shape sampling, an adaptive scheme for proposal variances, and a log-likelihood tempering strategy to estimate our model. Third, we validate our algorithm on 2D simulated data, and then estimate a scenario of alteration of the shape of the hippocampus 3D brain structure during the course of Alzheimer's disease. The method shows for instance that hippocampal atrophy progresses more quickly in female subjects, and occurs earlier in APOE4 mutation carriers. We finally illustrate the potential of our method for classifying pathological trajectories versus normal ageing.

****

CNN Driven Sparse Multi-Level B-Spline Image Registration

Pingge Jiang, James A. Shackleford; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9281-9289

Traditional single-grid and pyramidal B-spline parameterizations used in deformable image registration require users to specify control point spacing configurations capable of accurately capturing both global and complex local deformations. In many cases, such grid configurations are non-obvious and largely selected based on user experience. Recent regularization methods imposing sparsity upon the B-spline coefficients throughout simultaneous multi-grid optimization, however, have provided a promising means of determining suitable configurations automatically. Unfortunately, imposing sparsity on over-parameterized B-spline models is computationally expensive and introduces additional difficulties such as undesirable local minima in the B-spline coefficient optimization process. To over come these difficulties in determining B-spline grid configurations, this paper investigates the use of convolutional neural networks (CNNs) to learn and infer expressive sparse multi-grid configurations prior to B-spline coefficient optimization. Experimental results show that multi-grid configurations produced in this fashion using our CNN based approach provide registration quality comparable to L1-norm constrained over-parameterizations in terms of exactness, while exhibiting significantly reduced computational requirements.

****

Anatomical Priors in Convolutional Networks for Unsupervised Biomedical Segmentation

Adrian V. Dalca, John Guttag, Mert R. Sabuncu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9290-9299

We consider the problem of segmenting a biomedical image into anatomical regions of interest. We specifically address the frequent scenario where we have no paired training data that contains images and their manual segmentations. Instead, we employ unpaired segmentation images that we use to build an anatomical prior. Critically these segmentations can be derived from imaging data from a different dataset and imaging modality than the current task. We introduce a generative probabilistic model that employs the learned prior through a convolutional neural network to compute segmentations in an unsupervised setting. We conducted an empirical analysis of the proposed approach in the context of structural brain MRI segmentation, using a multi-study dataset of more than 14,000 scans. Our results show that an anatomical prior enables fast unsupervised segmentation which is typically not possible using standard convolutional networks. The integration of anatomical priors can facilitate CNN-based anatomical segmentation in a range of novel clinical problems, where few or no annotations are available and thus standard networks are not trainable. The code, model definitions and model weights are freely available at http://github.com/adalca/neuron

****

3D Registration of Curves and Surfaces Using Local Differential Information

Carolina Raposo, João P. Barreto; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9300-9308

This article presents for the first time a global method for registering 3D curves with 3D surfaces without requiring an initialization. The algorithm works with 2-tuples point+vector that consist in pairs of points augmented with the information of their tangents or normals. A closed-form solution for determining the alignment transformation from a pair of matching 2-tuples is proposed. In addition, the set of necessary conditions for two 2-tuples to match is derived. This allows fast search of correspondences that are used in an hypothesise-and-test framework for accomplishing global registration. Comparative experiments demonstrate that the proposed algorithm is the first effective solution for curve vs surface registration, with the method achieving accurate alignment in situations of small overlap and large percentage of outliers in a fraction of a second. The proposed framework is extended to the cases of curve vs curve and surface vs surface registration, with the former being particularly relevant since it is also a largely unsolved problem.

****

Weakly Supervised Learning of Single-Cell Feature Embeddings

Juan C. Caicedo, Claire McQuin, Allen Goodman, Shantanu Singh, Anne E. Carpenter; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9309-9318

Many new applications in drug discovery and functional genomics require capturing the morphology of individual imaged cells as comprehensively as possible rather than measuring one particular feature. In these so-called profiling experiments, the goal is to compare populations of cells treated with different chemicals or genetic perturbations in order to identify biomedically important similarities. Deep convolutional neural networks (CNNs) often make excellent feature extractors but require ground truth for training; this is rarely available in biomedical profiling experiments. We therefore propose to train CNNs based on a weakly supervised approach, where the network aims to classify each treatment against al

l others. Using this network as a feature extractor performed comparably to a network trained on non-biological, natural images on a chemical screen benchmark task, and improved results significantly on a more challenging genetic benchmark presented for the first time.
****
Guided Proofreading of Automatic Segmentations for Connectomics
Daniel Haehn, Verena Kaynig, James Tompkin, Jeff W. Lichtman, Hanspeter Pfister; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9319-9328
Automatic cell image segmentation methods in connectomics produce merge and split errors, which require correction through proofreading. Previous research has identified the visual search for these errors as the bottleneck in interactive proofreading. To aid error correction, we develop two classifiers that automatically recommend candidate merges and splits to the user. These classifiers use a convolutional neural network (CNN) that has been trained with errors in automatic segmentations against expert-labeled ground truth. Our classifiers detect potentially-erroneous regions by considering a large context region around a segmentation boundary. Corrections can then be performed by a user with yes/no decisions, which reduces variation of information 7.5x faster than previous proofreading methods. We also present a fully-automatic mode that uses a probability threshold to make merge/split decisions. Extensive experiments using the automatic approach and comparing performance of novice and expert users demonstrate that our method performs favorably against state-of-the-art proofreading methods on different connectomics datasets.
****
Wide Compression: Tensor Ring Nets
Wenqi Wang, Yifan Sun, Brian Eriksson, Wenlin Wang, Vaneet Aggarwal; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9329-9338
Deep neural networks have demonstrated state-of-the-art performance in a variety of real-world applications. In order to obtain performance gains, these networks have grown larger and deeper, containing millions or even billions of parameters and over a thousand layers. The trade-off is that these large architectures require an enormous amount of memory, storage, and computation, thus limiting their usability. Inspired by the recent tensor ring factorization, we introduce Tensor Ring Networks (TR-Nets), which significantly compress both the fully connected layers and the convolutional layers of deep networks. Our results show that our TR-Nets approach is able to compress LeNet-5 by 11x without losing accuracy, and can compress the state-of-the-art Wide ResNet by 243x with only 2.3% degradation in Cifar10 image classification. Overall, this compression scheme shows promise in scientific computing and deep learning, especially for emerging resource-constrained devices such as smartphones, wearables, and IoT devices.
****
Improvements to Context Based Self-Supervised Learning
T. Nathan Mundhenk, Daniel Ho, Barry Y. Chen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9339-9348
We develop a set of methods to improve on the results of self-supervised learning using context. We start with a baseline of patch based arrangement context learning and go from there. Our methods address some overt problems such as chromatic aberration as well as other potential problems such as spatial skew and mid-level feature neglect. We prevent problems with testing generalization on common self-supervised benchmark tests by using different datasets during our development. The results of our methods combined yield top scores on all standard self-supervised benchmarks, including classification and detection on PASCAL VOC 2007, segmentation on PASCAL VOC 2012, and "linear tests" on the ImageNet and CSAIL Places datasets. We obtain an improvement over our baseline method of between 4.0 to 7.1 percentage points on transfer learning classification tests. We also show results on different standard network architectures to demonstrate generalization as well as portability. All data, models and programs are available at: https://gdo-datasci. llnl.gov/selfsupervised/.

****

Learning Structure and Strength of CNN Filters for Small Sample Size Training

Rohit Keshari, Mayank Vatsa, Richa Singh, Afzel Noore; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9349-9358

Convolutional Neural Networks have provided state-of-the-art results in several computer vision problems. However, due to a large number of parameters in CNNs, they require a large number of training samples which is a limiting factor for small sample size problems. To address this limitation, in this paper, we propose SSF-CNN which focuses on learning the “structure" and “strength" of filters. The structure of the filter is initialized using a dictionary based filter learning algorithm and the strength of the filter is learned using the small sample training data. The architecture provides the flexibility of training with both small and large training databases, and yields good accuracies even with small size training data. The effectiveness of the algorithm is demonstrated on MNIST, CIFAR10, NORB, Omniglot, and Newborn Face Image databases, with varying number of training samples. The results show that SSF-CNN significantly reduces the number of parameters required for training while providing high accuracies on the test database. On small problems such as newborn face recognition, the results demonstrate improvement in rank-1 identification accuracy by at least 10%.

****

Boosting Self-Supervised Learning via Knowledge Transfer

Mehdi Noroozi, Ananth Vinjimoor, Paolo Favaro, Hamed Pirsiavash; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9359-9367

In self-supervised learning one trains a model to solve a so-called pretext task on a dataset without the need for human annotation. The main objective, however, is to transfer this model to a target domain and task. Currently, the most effective transfer strategy is fine-tuning, which restricts one to use the same model or parts thereof for both pretext and target tasks. In this paper, we present a novel framework for self-supervised learning that overcomes limitations in designing and comparing different tasks, models, and data domains. In particular, our framework decouples the structure of the self-supervised model from the final task-specific fine-tuned model. This allows us to: 1) quantitatively assess previously incompatible models including handcrafted features; 2) show that deeper neural network models can learn better representations from the same pretext task; 3) transfer knowledge learned with a deep model to a shallower one and thus boost its learning. We use this framework to design a novel self-supervised task, which achieves state-of-the-art performance on the common benchmarks in PASCAL VOC 2007, ILSVRC12 and Places by a significant margin. A surprising result is that our learned features shrink the mAP gap between models trained via self-supervised learning and supervised learning from 5.9 to 2.6 in object detection on PASCAL VOC 2007.

****

The Power of Ensembles for Active Learning in Image Classification

William H. Beluch, Tim Genewein, Andreas Nürnberger, Jan M. Köhler; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9368-9377

Deep learning methods have become the de-facto standard for challenging image processing tasks such as image classification. One major hurdle of deep learning approaches is that large sets of labeled data are necessary, which can be prohibitively costly to obtain, particularly in medical image diagnosis applications. Active learning techniques can alleviate this labeling effort. In this paper we investigate some recently proposed methods for active learning with high-dimensional data and convolutional neural network classifiers. We compare ensemble-based methods against Monte-Carlo Dropout and geometric approaches. We find that ensembles perform better and lead to more calibrated predictive uncertainties, which are the basis for many active learning algorithms. To investigate why Monte-Carlo Dropout uncertainties perform worse, we explore potential differences in isolation in a series of experiments. We show results for MNIST and CIFAR-10, on which we achieve a test set accuracy of $90 %$ with roughly 12,200 labeled images, and initial results on ImageNet. Additionally, we show results on a large, highly class-imbalanced diabetic retinopathy dataset. We observe that the ensemble-based active learning effectively counteracts this imbalance during acquisition.

****

Learning Compact Recurrent Neural Networks With Block-Term Tensor Decomposition

Jinmian Ye, Linnan Wang, Guangxi Li, Di Chen, Shandian Zhe, Xinqi Chu, Zenglin Xu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9378-9387

Recurrent Neural Networks (RNNs) are powerful sequence modeling tools. However, when dealing with high dimensional inputs, the training of RNNs becomes computational expensive due to the large number of model parameters. This hinders RNNs from solving many important computer vision tasks, such as Action Recognition in Videos and Image Captioning. To overcome this problem, we propose a compact and flexible structure, namely Block-Term tensor decomposition, which greatly reduces the parameters of RNNs and improves their training efficiency. Compared with alternative low-rank approximations, such as tensor-train RNN (TT-RNN), our method, Block-Term RNN (BT-RNN), is not only more concise (when using the same rank), but also able to attain a better approximation to the original RNNs with much fewer parameters. On three challenging tasks, including Action Recognition in Videos, Image Captioning and Image Generation, BT-RNN outperforms TT-RNN and the standard RNN in terms of both prediction accuracy and convergence rate. Specifically, BT-LSTM utilizes 17,388 times fewer parameters than the standard LSTM to achieve an accuracy improvement over 15.6% in the Action Recognition task on the UCF11 dataset.

****

Spatially-Adaptive Filter Units for Deep Neural Networks

Domen Tabernik, Matej Kristan, Aleš Leonardis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9388-9396

Classical deep convolutional networks increase receptive field size by either gradual resolution reduction or application of hand-crafted dilated convolutions to prevent increase in the number of parameters. In this paper we propose a novel displaced aggregation unit (DAU) that does not require hand-crafting. In contrast to classical filters with units (pixels) placed on a fixed regular grid, the displacement of the DAUs are learned, which enables filters to spatially-adapt their receptive field to a given problem. We extensively demonstrate the strength of DAUs on a classification and semantic segmentation tasks. Compared to ConvNets with regular filter, ConvNets with DAUs achieve comparable performance at faster convergence and up to 3-times reduction in parameters. Furthermore, DAUs allow us to study deep networks from novel perspectives. We study spatial distributions of DAU filters and analyze the number of parameters allocated for spatial coverage in a filter.

****

SO-Net: Self-Organizing Network for Point Cloud Analysis

Jiaxin Li, Ben M. Chen, Gim Hee Lee; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9397-9406

This paper presents SO-Net, a permutation invariant architecture for deep learning with orderless point clouds. The SO-Net models the spatial distribution of point cloud by building a Self-Organizing Map (SOM). Based on the SOM, SO-Net performs hierarchical feature extraction on individual points and SOM nodes, and ultimately represents the input point cloud by a single feature vector. The receptive field of the network can be systematically adjusted by conducting point-to-node k nearest neighbor search. In recognition tasks such as point cloud reconstruction, classification, object part segmentation and shape retrieval, our proposed network demonstrates performance that is similar with or better than state-of-the-art approaches. In addition, the training speed is significantly faster than existing point cloud recognition networks because of the parallelizability and simplicity of the proposed architecture. Our code is available at the project website.

****

SGAN: An Alternative Training of Generative Adversarial Networks

Tatjana Chavdarova, François Fleuret; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9407-9415

The Generative Adversarial Networks (GANs) have demonstrated impressive performance for data synthesis, and are now used in a wide range of computer vision tasks. In spite of this success, they gained a reputation for being difficult to train, what results in a time-consuming and human-involved development process to use them. We consider an alternative training process, named SGAN, in which several adversarial "local" pairs of networks are trained independently so that a "global" supervising pair of networks can be trained against them. The goal is to train the global pair with the corresponding ensemble opponent for improved performances in terms of mode coverage. This approach aims at increasing the chances that learning will not stop for the global pair, preventing both to be trapped in an unsatisfactory local minimum, or to face oscillations often observed in practice. To guarantee the latter, the global pair never affects the local ones. The rules of SGAN training are thus as follows: the global generator and discriminator are trained using the local discriminators and generators, respectively, whereas the local networks are trained with their fixed local opponent. Experimental results on both toy and real-world problems demonstrate that this approach outperforms standard training in terms of better mitigating mode collapse, stability while converging and that it surprisingly, increases the convergence speed as well.

****

SketchyGAN: Towards Diverse and Realistic Sketch to Image Synthesis

Wengling Chen, James Hays; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9416-9425

Synthesizing realistic images from human drawn sketches is a challenging problem in computer graphics and vision. Existing approaches either need exact edge maps, or rely on retrieval of existing photographs. In this work, we propose a novel Generative Adversarial Network (GAN) approach that synthesizes plausible images from 50 categories including motorcycles, horses and couches. We demonstrate a data augmentation technique for sketches which is fully automatic, and we show that the augmented data is helpful to our task. We introduce a new network building block suitable for both the generator and discriminator which improves the information flow by injecting the input image at multiple scales. Compared to state-of-the-art image translation methods, our approach generates more realistic images and achieves significantly higher Inception Scores.

****

Explicit Loss-Error-Aware Quantization for Low-Bit Deep Neural Networks

Aojun Zhou, Anbang Yao, Kuan Wang, Yurong Chen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9426-9435

Benefiting from tens of millions of hierarchically stacked learnable parameters, Deep Neural Networks (DNNs) have demonstrated overwhelming accuracy on a variety of artificial intelligence tasks. However reversely, the large size of DNN models lays a heavy burden on storage, computation and power consumption, which prohibits their deployments on the embedded and mobile systems. In this paper, we propose Explicit Loss-error-aware Quantization (ELQ), a new method that can train DNN models with very low-bit parameter values such as ternary and binary ones to approximate 32-bit floating-point counterparts without noticeable loss of predication accuracy. Unlike existing methods that usually pose the problem as a straightforward approximation of the layer-wise weights or outputs of the original full-precision model (specifically, minimizing the error of the layer-wise weights or inner products of the weights and the inputs between the original and respective quantized models), our ELQ elaborately bridges the loss perturbation from the weight quantization and an incremental quantization strategy to address DNN quantization. Through explicitly regularizing the loss perturbation and the weight approximation error in an incremental way, we show that such a new optimization method is theoretically reasonable and practically effective. As validated with two mainstream convolutional neural network families (i.e., fully convolutional and non-fully convolutional), our ELQ shows better results than the state-of-the-art quantization methods on the large scale ImageNet classification dataset

. Code will be made publicly available.

****

Towards Universal Representation for Unseen Action Recognition

Yi Zhu, Yang Long, Yu Guan, Shawn Newsam, Ling Shao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9436-9445

Unseen Action Recognition (UAR) aims to recognise novel action categories without training examples. While previous methods focus on inner-dataset seen/unseen splits, this paper proposes a pipeline using a large-scale training source to achieve a Universal Representation (UR) that can generalise to a more realistic Cross-Dataset UAR (CD-UAR) scenario. We first address UAR as a Generalised Multiple-Instance Learning (GMIL) problem and discover "building-blocks" from the large-scale ActivityNet dataset using distribution kernels. Essential visual and semantic components are preserved in a shared space to achieve the UR that can efficiently generalise to new datasets. Predicted UR exemplars can be improved by a simple semantic adaptation, and then an unseen action can be directly recognised using UR during the test. Without further training, extensive experiments manifest significant improvements over the UCF101 and HMDB51 benchmarks.

****

Deep Image Prior

Dmitry Ulyanov, Andrea Vedaldi, Victor Lempitsky; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9446-9454

Deep convolutional networks have become a popular tool for image generation and restoration. Generally, their excellent performance is imputed to their ability to learn realistic image priors from a large number of example images. In this paper, we show that, on the contrary, the structure of a generator network is sufficient to capture a great deal of low-level image statistics prior to any learning. In order to do so, we show that a randomly-initialized neural network can be used as a handcrafted prior with excellent results in standard inverse problems such as denoising, super-resolution, and inpainting. Furthermore, the same prior can be used to invert deep neural representations to diagnose them, and to restore images based on flash-no flash input pairs. Apart from its diverse applications, our approach highlights the inductive bias captured by standard generator network architectures. It also bridges the gap between two very popular families of image restoration methods: learning-based methods using deep convolutional networks and learning-free methods based on handcrafted image priors such as self-similarity.

****

ST-GAN: Spatial Transformer Generative Adversarial Networks for Image Compositing

Chen-Hsuan Lin, Ersin Yumer, Oliver Wang, Eli Shechtman, Simon Lucey; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9455-9464

We address the problem of finding realistic geometric corrections to a foreground object such that it appears natural when composited into a background image. To achieve this, we propose a novel Generative Adversarial Network (GAN) architecture that utilizes Spatial Transformer Networks (STNs) as the generator, which we call Spatial Transformer GANs (ST-GANs). ST-GANs seek image realism by operating in the geometric warp parameter space. In particular, we exploit an iterative STN warping scheme and propose a sequential training strategy that achieves better results compared to naive training of a single generator. One of the key advantages of ST-GAN is its applicability to high-resolution images indirectly since the predicted warp parameters are transferable between reference frames. We demonstrate our approach in two applications: (1) visualizing how indoor furniture (e.g. from product images) might be perceived in a room, (2) hallucinating how accessories like glasses would look when matched with real portraits.

****

CartoonGAN: Generative Adversarial Networks for Photo Cartoonization

Yang Chen, Yu-Kun Lai, Yong-Jin Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9465-9474

In this paper, we propose a solution to transforming photos of real-world scenes into cartoon style images, which is valuable and challenging in computer vision and computer graphics. Our solution belongs to learning based methods, which have recently become popular to stylize images in artistic forms such as painting. However, existing methods do not produce satisfactory results for cartoonization, due to the fact that (1) cartoon styles have unique characteristics with high level simplification and abstraction, and (2) cartoon images tend to have clear edges, smooth color shading and relatively simple textures, which exhibit significant challenges for texture-descriptor-based loss functions used in existing methods. In this paper, we propose CartoonGAN, a generative adversarial network (GAN) framework for cartoon stylization. Our method takes unpaired photos and cartoon images for training, which is easy to use. Two novel losses suitable for cartoonization are proposed: (1) a semantic content loss, which is formulated as a sparse regularization in the high-level feature maps of the VGG network to cope with substantial style variation between photos and cartoons, and (2) an edge-promoting adversarial loss for preserving clear edges. We further introduce an initialization phase, to improve the convergence of the network to the target manifold. Our method is also much more efficient to train than existing methods. Experimental results show that our method is able to generate high-quality cartoon images from real-world photos (i.e., following specific artists' styles and with clear edges and smooth shading) and outperforms state-of-the-art methods.

****