Embodied Question Answering

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, Dhruv Batra; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1-10

We present a new AI task -- Embodied Question Answering (EmbodiedQA) -- where an agent is spawned at a random location in a 3D environment and asked a question ("What color is the car?"). In order to answer, the agent must first intelligently navigate to explore the environment, gather necessary visual information through first-person (egocentric) vision, and then answer the question ("orange"). EmbodiedQA requires a range of AI skills -- language understanding, visual recognition, active perception, goal-driven navigation, commonsense reasoning, long-term memory, and grounding language into actions. In this work, we develop a dataset of questions and answers in House3D environments, evaluation metrics, and a hierarchical model trained with imitation and reinforcement learning.

****

Learning by Asking Questions

Ishan Misra, Ross Girshick, Rob Fergus, Martial Hebert, Abhinav Gupta, Laurens van der Maaten; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 11-20

We introduce an interactive learning framework for the development and testing of intelligent visual systems, called learning-by-asking (LBA). We explore LBA in context of the Visual Question Answering (VQA) task. LBA differs from standard VQA training in that most questions are not observed during training time, and the learner must ask questions it wants answers to. Thus, LBA more closely mimics natural learning and has the potential to be more data-efficient than the traditional VQA setting. We present a model that performs LBA on the CLEVR dataset, and show that it automatically discovers an easy-to-hard curriculum when learning interactively from an oracle. Our LBA generated data consistently matches or outperforms the CLEVR train data and is more sample efficient. We also show that our model asks questions that generalize to state-of-the-art VQA models and to novel test time distributions.

****

Finding Tiny Faces in the Wild With Generative Adversarial Network

Yancheng Bai, Yongqiang Zhang, Mingli Ding, Bernard Ghanem; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 21-30

Face detection techniques have been developed for decades, and one of remaining open challenges is detecting small faces in unconstrained conditions. The reason is that tiny faces are often lacking detailed information and blurring. In this paper, we proposed an algorithm to directly generate a clear high-resolution face from a blurry small one by adopting a generative adversarial network (GAN). Toward this end, the basic GAN formulation achieves it by super-resolving and refining sequentially (e.g. SR-GAN and cycle-GAN). However, we design a novel network to address the problem of super-resolving and refining jointly. We also introduce new training losses to guide the generator network to recover fine details and to promote the discriminator network to distinguish real vs. fake and face vs. non-face simultaneously. Extensive experiments on the challenging dataset WIDER FACE demonstrate the effectiveness of our proposed method in restoring a clear high-resolution face from a blurry small one, and show that the detection performance outperforms other state-of-the-art methods.

****

Learning Face Age Progression: A Pyramid Architecture of GANs

Hongyu Yang, Di Huang, Yunhong Wang, Anil K. Jain; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 31-39

The two underlying requirements of face age progression, i.e. aging accuracy and identity permanence, are not well studied in the literature. In this paper, we present a novel generative adversarial network based approach. It separately models the constraints for the intrinsic subject-specific characteristics and the age-specific facial changes with respect to the elapsed time, ensuring that the generated faces present desired aging effects while simultaneously keeping person alized properties stable. Further, to generate more lifelike facial details, high-level age-specific features conveyed by the synthesized face are estimated by a pyramidal adversarial discriminator at multiple scales, which simulates the aging effects in a finer manner. The proposed method is applicable to diverse face samples in the presence of variations in pose, expression, makeup, etc., and remarkably vivid aging effects are achieved. Both visual fidelity and quantitative evaluations show that the approach advances the state-of-the-art.

****

PairedCycleGAN: Asymmetric Style Transfer for Applying and Removing Makeup
Huiwen Chang, Jingwan Lu, Fisher Yu, Adam Finkelstein; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 40-48

This paper introduces an automatic method for editing a portrait photo so that the subject appears to be wearing makeup in the style of another person in a reference photo. Our unsupervised learning approach relies on a new framework of cycle-consistent generative adversarial networks. Different from the image domain transfer problem, our style transfer problem involves two asymmetric functions: a forward function encodes example-based style transfer, whereas a backward function removes the style. We construct two coupled networks to implement these functions -- one that transfers makeup style and a second that can remove makeup -- such that the output of their successive application to an input photo will match the input. The learned style network can then quickly apply an arbitrary makeup style to an arbitrary photo. We demonstrate the effectiveness on a broad range of portraits and styles.

****

GANerated Hands for Real-Time 3D Hand Tracking From Monocular RGB
Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, Christian Theobalt; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 49-59

We address the highly challenging problem of real-time 3D hand tracking based on a monocular RGB-only sequence. Our tracking method combines a convolutional neural network with a kinematic 3D hand model, such that it generalizes well to unseen data, is robust to occlusions and varying camera viewpoints, and leads to anatomically plausible as well as temporally smooth hand motions. For training our CNN we propose a novel approach for the synthetic generation of training data that is based on a geometrically consistent image-to-image translation network. To be more specific, we use a neural network that translates synthetic images to "real" images, such that the so-generated images follow the same statistical distribution as real-world hand images. For training this translation network we combine an adversarial loss and a cycle-consistency loss with a geometric consistency loss in order to preserve geometric properties (such as hand pose) during translation. We demonstrate that our hand tracking system outperforms the current state-of-the-art on challenging RGB-only footage.

****

Learning Pose Specific Representations by Predicting Different Views
Georg Poier, David Schinagl, Horst Bischof; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 60-69

The labeled data required to learn pose estimation for articulated objects is difficult to provide in the desired quantity, realism, density, and accuracy. To address this issue, we develop a method to learn representations, which are very specific for articulated poses, without the need for labeled training data. We exploit the observation that the object pose of a known object is predictive for the appearance in any known view. That is, given only the pose and shape parameters of a hand, the hand's appearance from any viewpoint can be approximated. To exploit this observation, we train a model that - given input from one view - estimates a latent representation, which is trained to be predictive for the appearance of the object when captured from another viewpoint. Thus, the only necessary supervision is the second view. The training process of this model reveals an implicit pose representation in the latent space. Importantly, at test time the pose representation can be inferred using only a single view. In qualitative and quantitative experiments we show that the learned representations capture deta iled pose information. Moreover, when training the proposed method jointly with labeled and unlabeled data, it consistently surpasses the performance of its fully supervised counterpart, while reducing the amount of needed labeled samples by at least one order of magnitude.

****

Weakly and Semi Supervised Human Body Part Parsing via Pose-Guided Knowledge Transfer

Hao-Shu Fang, Guansong Lu, Xiaolin Fang, Jianwen Xie, Yu-Wing Tai, Cewu Lu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 70-78

Human body part parsing, or human semantic part segmentation, is fundamental to many computer vision tasks. In conventional semantic segmentation methods, the ground truth segmentations are provided, and fully convolutional networks (FCN) are trained in an end-to-end scheme. Although these methods have demonstrated impressive results, their performance highly depends on the quantity and quality of training data. In this paper, we present a novel method to generate synthetic human part segmentation data using easily-obtained human keypoint annotations. Our key idea is to exploit the anatomical similarity among human to transfer the parsing results of a person to another person with similar pose. Using these estimated results as additional training data, our semi-supervised model outperforms its strong-supervised counterpart by 6 mIOU on the PASCAL-Person-Part dataset, and we achieve state-of-the-art human parsing results. Our approach is general and can be readily extended to other object/animal parsing task assuming that their anatomical similarity can be annotated by keypoints. The proposed model and accompanying source code will be made publicly available.

****

Person Transfer GAN to Bridge Domain Gap for Person Re-Identification

Longhui Wei, Shiliang Zhang, Wen Gao, Qi Tian; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 79-88

Although the performance of person Re-Identification (ReID) has been significantly boosted, many challenging issues in real scenarios have not been fully investigated, e.g., the complex scenes and lighting variations, viewpoint and pose changes, and the large number of identities in a camera network. To facilitate the research towards conquering those issues, this paper contributes a new dataset called MSMT17 with many important features, e.g., 1) the raw videos are taken by an 15-camera network deployed in both indoor and outdoor scenes, 2) the videos cover a long period of time and present complex lighting variations, and 3) it contains currently the largest number of annotated identities, i.e., 4,101 identities and 126,441 bounding boxes. We also observe that, domain gap commonly exists between datasets, which essentially causes severe performance drop when training and testing on different datasets. This results in that available training data cannot be effectively leveraged for new testing domains. To relieve the expensive costs of annotating new training samples, we propose a Person Transfer Generative Adversarial Network (PTGAN) to bridge the domain gap. Comprehensive experiments show that the domain gap could be substantially narrowed-down by the PTGAN.

****

Cross-Modal Deep Variational Hand Pose Estimation

Adrian Spurr, Jie Song, Seonwook Park, Otmar Hilliges; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 89-98

The human hand moves in complex and high-dimensional ways, making estimation of 3D hand pose configurations from images alone a challenging task. In this work we propose a method to learn a statistical hand model represented by a cross-modal trained latent space via a generative deep neural network. We derive an objective function from the variational lower bound of the VAE framework and jointly optimize the resulting cross-modal KL-divergence and the posterior reconstruction objective, naturally admitting a training regime that leads to a coherent latent space across multiple modalities such as RGB images, 2D keypoint detections or 3D hand configurations. Additionally, it grants a straightforward way of using semi-supervision. This latent space can be directly used to estimate 3D hand pos es from RGB images, outperforming the state-of-the art in different settings. Furthermore, we show that our proposed method can be used without changes on depth images and performs comparably to specialized methods. Finally, the model is fully generative and can synthesize consistent pairs of hand configurations across modalities. We evaluate our method on both RGB and depth datasets and analyze the latent space qualitatively.

****

Disentangled Person Image Generation

Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, Mario Fritz; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 99-108

Generating novel, yet realistic, images of persons is a challenging task due to the complex interplay between the different image factors, such as the foreground, background and pose information. In this work, we aim at generating such images based on a novel, two-stage reconstruction pipeline that learns a disentangled representation of the aforementioned image factors and generates novel person images at the same time. First, a multi-branched reconstruction network is proposed to disentangle and encode the three factors into embedding features, which are then combined to re-compose the input image itself. Second, three corresponding mapping functions are learned in an adversarial manner in order to map Gaussian noise to the learned embedding feature space, for each factor, respectively. Using the proposed framework, we can manipulate the foreground, background and pose of the input image, and also sample new embedding features to generate such targeted manipulations, that provide more control over the generation process. Experiments on the Market-1501 and Deepfashion datasets show that our model does not only generate realistic person images with new foregrounds, backgrounds and poses, but also manipulates the generated factors and interpolates the in-between states. Another set of experiments on Market-1501 shows that our model can also be beneficial for the person re-identification task.

****

Super-FAN: Integrated Facial Landmark Localization and Super-Resolution of Real-World Low Resolution Faces in Arbitrary Poses With GANs

Adrian Bulat, Georgios Tzimiropoulos; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 109-117

This paper addresses 2 challenging tasks: improving the quality of low resolution facial images and accurately locating the facial landmarks on such poor resolution images. To this end, we make the following 5 contributions: (a) we propose Super-FAN: the very first end-to-end system that addresses both tasks simultaneously, i.e. both improves face resolution and detects the facial landmarks. The novelty or Super-FAN lies in incorporating structural information in a GAN-based super-resolution algorithm via integrating a sub-network for face alignment through heatmap regression and optimizing a novel heatmap loss. (b) We illustrate the benefit of training the two networks jointly by reporting good results not only on frontal images (as in prior work) but on the whole spectrum of facial poses, and not only on synthetic low resolution images (as in prior work) but also on real-world images. (c) We improve upon the state-of-the-art in face super-resolution by proposing a new residual-based architecture. (d) Quantitatively, we show large improvement over the state-of-the-art for both face super-resolution and alignment. (e) Qualitatively, we show for the first time good results on real-world low resolution images.

****

Multistage Adversarial Losses for Pose-Based Human Image Synthesis

Chenyang Si, Wei Wang, Liang Wang, Tieniu Tan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 118-126

Human image synthesis has extensive practical applications e.g. person re-identification and data augmentation for human pose estimation. However, it is much more challenging than rigid object synthesis, e.g. cars and chairs, due to the variability of human posture. In this paper, we propose a pose-based human image synthesis method which can keep the human posture unchanged in novel viewpoints. Furthermore, we adopt multistage adversarial losses separately for the foreground and background generation, which fully exploits the multi-modal characteristics of generative loss to generate more realistic looking images. We perform extensive experiments on the Human3.6M dataset and verify the effectiveness of each stage of our method. The generated human images not only keep the same pose as the input image, but also have clear detailed foreground and background. The quantitative comparison results illustrate that our approach achieves much better results than several state-of-the-art methods.

****

Rotation Averaging and Strong Duality

Anders Eriksson, Carl Olsson, Fredrik Kahl, Tat-Jun Chin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 127-135

In this paper we explore the role of duality principles within the problem of rotation averaging, a fundamental task in a wide range of computer vision applications. In its conventional form, rotation averaging is stated as a minimization over multiple rotation constraints. As these constraints are non-convex, this problem is generally considered challenging to solve globally. We show how to circumvent this difficulty through the use of Lagrangian duality. While such an approach is well-known it is normally not guaranteed to provide a tight relaxation. Based on spectral graph theory, we analytically prove that in many cases there is no duality gap unless the noise levels are severe. This allows us to obtain certifiably global solutions to a class of important non-convex problems in polynomial time.   We also propose an efficient, scalable algorithm that out-performs general purpose numerical solvers and is able to handle the large problem instances commonly occurring in structure from motion settings. The potential of this proposed method is demonstrated on a number of different problems, consisting of both synthetic and real-world data.

****

Hybrid Camera Pose Estimation

Federico Camposeco, Andrea Cohen, Marc Pollefeys, Torsten Sattler; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 136-144

In this paper, we aim to solve the pose estimation problem of calibrated pinhole and generalized cameras w.r.t. a Structure-from-Motion (SfM) model by leveraging both 2D-3D correspondences as well as 2D-2D correspondences. Traditional approaches either focus on the use of 2D-3D matches, known as structure-based pose estimation or solely on 2D-2D matches (structure-less pose estimation). Absolute pose approaches are limited in their performance by the quality of the 3D point triangulations as well as the completeness of the 3D model. Relative pose approaches, on the other hand, while being more accurate, also tend to be far more computationally costly and often return dozens of possible solutions. This work aims to bridge the gap between these two paradigms. We propose a new RANSAC-based approach that automatically chooses the best type of solver to use at each iteration in a data-driven way. The solvers chosen by our RANSAC can range from pure structure-based or structure-less solvers, to any possible combination of hybrid solvers (i.e. using both types of matches) in between. A number of these new hybrid minimal solvers are also presented in this paper. Both synthetic and real data experiments show our approach to be as accurate as structure-less approaches, while staying close to the efficiency of structure-based methods.

****

A Certifiably Globally Optimal Solution to the Non-Minimal Relative Pose Problem

Jesus Briales, Laurent Kneip, Javier Gonzalez-Jimenez; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 145-154

Finding the relative pose between two calibrated views ranks among the most fundamental geometric vision problems. It therefore appears as somewhat a surprise that a globally optimal solver that minimizes a properly defined energy over non-minimal correspondence sets and in the original space of relative transformations has yet to be discovered. This, notably, is the contribution of the present paper. We formulate the problem as a Quadratically Constrained Quadratic Program (QCQP), which can be converted into a Semidefinite Program (SDP) using Shor's con vex relaxation. While a theoretical proof for the tightness of this relaxation remains open, we prove through exhaustive validation on both simulated and real experiments that our approach always finds and certifies (a-posteriori) the global optimum of the cost function.

****

Single View Stereo Matching

Yue Luo, Jimmy Ren, Mude Lin, Jiahao Pang, Wenxiu Sun, Hongsheng Li, Liang Lin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 155-163

Previous monocular depth estimation methods take a single view and directly regress the expected results. Though recent advances are made by applying geometrically inspired loss functions during training, the inference procedure does not explicitly impose any geometrical constraint. Therefore these models purely rely on the quality of data and the effectiveness of learning to generalize. This either leads to suboptimal results or the demand of huge amount of expensive ground truth labelled data to generate reasonable results. In this paper, we show for the first time that the monocular depth estimation problem can be reformulated as two sub-problems, a view synthesis procedure followed by stereo matching, with two intriguing properties, namely i) geometrical constraints can be explicitly imposed during inference; ii) demand on labelled depth data can be greatly alleviated. We show that the whole pipeline can still be trained in an end-to-end fashion and this new formulation plays a critical role in advancing the performance. The resulting model outperforms all the previous monocular depth estimation methods as well as the stereo block matching method in the challenging KITTI dataset by only using a small number of real training data. The model also generalizes well to other monocular depth estimation benchmarks. We also discuss the implications and the advantages of solving monocular depth estimation using stereo methods.

****

Fight Ill-Posedness With Ill-Posedness: Single-Shot Variational Depth Super-Resolution From Shading

Bjoern Haefner, Yvain Quéau, Thomas Möllenhoff, Daniel Cremers; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 164-174

We put forward a principled variational approach for up-sampling a single depth map to the resolution of the companion color image provided by an RGB-D sensor. We combine heterogeneous depth and color data in order to jointly solve the ill-posed depth super-resolution and shape-from-shading problems. The low-frequency geometric information necessary to disambiguate shape-from-shading is extracted from the low-resolution depth measurements and, symmetrically, the high-resolution photometric clues in the RGB image provide the high-frequency information required to disambiguate depth super-resolution.

****

Deep Depth Completion of a Single RGB-D Image

Yinda Zhang, Thomas Funkhouser; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 175-185

The goal of our work is to complete the depth channel of an RGB-D image. Commodity-grade depth cameras often fail to sense depth for shiny, bright, transparent, and distant surfaces. To address this problem, we train a deep network that takes an RGB image as input and predicts dense surface normals and occlusion boundaries. Those predictions are then combined with raw depth observations provided by the RGB-D camera to solve for depths for all pixels, including those missing in the original observation. This method was chosen over others (e.g., inpainting depths directly) as the result of extensive experiments with a new depth completion benchmark dataset, where holes are filled in training data through the rendering of surface reconstructions created from multiview RGB-D scans. Experiments with different network inputs, depth representations, loss functions, optimization methods, inpainting methods, and deep depth estimation networks show that our proposed approach provides better depth completions than these alternatives.

****

Multi-View Harmonized Bilinear Network for 3D Object Recognition

Tan Yu, Jingjing Meng, Junsong Yuan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 186-194

View-based methods have achieved considerable success in 3D object recognition tasks. Different from existing view-based methods pooling the view-wise features, we tackle this problem from the perspective of patches-to-patches similarity measurement. By exploiting the relationship between polynomial kernel and bilinear pooling, we obtain an effective 3D object representation by aggregating local convolutional features through bilinear pooling. Meanwhile, we harmonize different components inherited in the pooled bilinear feature to obtain a more discriminative representation for a 3D object. To achieve an end-to-end trainable framework, we incorporate the harmonized bilinear pooling operation as a layer of a network, constituting the proposed Multi-view Harmonized Bilinear Network (MHBN). Systematic experiments conducted on two public benchmark datasets demonstrate the efficacy of the proposed methods in 3D object recognition.

****

PPFNet: Global Context Aware Local Features for Robust 3D Point Matching

Haowen Deng, Tolga Birdal, Slobodan Ilic; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 195-205

We present PPFNet - Point Pair Feature NETwork for deeply learning a globally informed 3D local feature descriptor to find correspondences in unorganized point clouds. PPFNet learns local descriptors on pure geometry and is highly aware of the global context, an important cue in deep learning. Our 3D representation is computed as a collection of point-pair-features combined with the points and normals within a local vicinity. Our permutation invariant network design is inspired by PointNet and sets PPFNet to be ordering-free. As opposed to voxelization, our method is able to consume raw point clouds to exploit the full sparsity. PPFNet uses a novel N-tuple loss and architecture injecting the global information naturally into the local descriptor. It shows that context awareness also boosts the local feature representation. Qualitative and quantitative evaluations of our network suggest increased recall, improved robustness and invariance as well as a vital step in the 3D descriptor extraction performance.

****

FoldingNet: Point Cloud Auto-Encoder via Deep Grid Deformation

Yaoqing Yang, Chen Feng, Yiru Shen, Dong Tian; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 206-215

Recent deep networks that directly handle points in a point set, e.g., PointNet, have been state-of-the-art for supervised learning tasks on point clouds such as classification and segmentation. In this work, a novel end-to-end deep auto-encoder is proposed to address unsupervised learning challenges on point clouds. On the encoder side, a graph-based enhancement is enforced to promote local structures on top of PointNet. Then, a novel folding-based decoder deforms a canonical 2D grid onto the underlying 3D object surface of a point cloud, achieving low reconstruction errors even for objects with delicate structures. The proposed decoder only uses about 7% parameters of a decoder with fully-connected neural networks, yet leads to a more discriminative representation that achieves higher linear SVM classification accuracy than the benchmark. In addition, the proposed decoder structure is shown, in theory, to be a generic architecture that is able to reconstruct an arbitrary point cloud from a 2D grid. Our code is available at http://www.merl.com/research/license#FoldingNet

****

A Papier-Mâché Approach to Learning 3D Surface Generation

Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan C. Russell, Mathieu Aubry; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 216-224

We introduce a method for learning to generate the surface of 3D shapes. Our approach represents a 3D shape as a collection of parametric surface elements and, in contrast to methods generating voxel grids or point clouds, naturally infers a surface representation of the shape. Beyond its novelty, our new shape generation framework, AtlasNet, comes with significant advantages, such as improved pre cision and generalization capabilities, and the possibility to generate a shape of arbitrary resolution without memory issues. We demonstrate these benefits and compare to strong baselines on the ShapeNet benchmark for two applications: (i) auto-encoding shapes, and (ii) single-view reconstruction from a still image. We also provide results showing its potentialfor other applications, such as morphing, parametrization, super-resolution, matching, and co-segmentation.

****

LEGO: Learning Edge With Geometry All at Once by Watching Videos

Zhenheng Yang, Peng Wang, Yang Wang, Wei Xu, Ram Nevatia; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 225-234

Learning to estimate 3D geometry in a single image by watching unlabeled videos via deep convolutional network is attracting significant attention. In this paper, we introduce a “3D as-smooth-as-possible (3D-ASAP)” prior inside the pipeline, which enables joint estimation of edges and 3D scene, yielding results with significant improvement in accuracy for fine detailed structures. Specifically, we define the 3D-ASAP prior by requiring that any two points recovered in 3D from an image should lie on an existing planar surface if no other cues provided. We design an unsupervised framework that Learns Edges and Geometry (depth, normal) all at Once (LEGO). The predicted edges are embedded into depth and surface normal smoothness terms, where pixels without edges in-between are constrained to satisfy the prior. In our framework, the predicted depths, normals and edges are forced to be consistent all the time. We conduct experiments on KITTI to evaluate our estimated geometry and CityScapes to perform edge evaluation. We show that in all of the tasks, i.e. depth, normal and edge, our algorithm vastly outperforms other state-of-the-art (SOTA) algorithms, demonstrating the benefits of our approach.

****

Five-Point Fundamental Matrix Estimation for Uncalibrated Cameras

Daniel Barath; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 235-243

We aim at estimating the fundamental matrix in two views from five correspondences of rotation invariant features obtained by e.g. the SIFT detector. The proposed minimal solver first estimates a homography from three correspondences assuming that they are co-planar and exploiting their rotational components. Then the fundamental matrix is obtained from the homography and two additional point pairs in general position. The proposed approach, combined with robust estimators like Graph-Cut RANSAC, is superior to other state-of-the-art algorithms both in terms of accuracy and number of iterations required. This is validated on synthesized data and 561 real image pairs. Moreover, the tests show that requiring three points on a plane is not too restrictive in urban environment and locally optimized robust estimators lead to accurate estimates even if the points are not entirely co-planar. As a potential application, we show that using the proposed method makes two-view multi-motion estimation more accurate.

****

PointFusion: Deep Sensor Fusion for 3D Bounding Box Estimation

Danfei Xu, Dragomir Anguelov, Ashesh Jain; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 244-253

We present PointFusion, a generic 3D object detection method that leverages both image and 3D point cloud information. Unlike existing methods that either use multi-stage pipelines or hold sensor and dataset-specific assumptions, PointFusion is conceptually simple and application-agnostic. The image data and the raw point cloud data are independently processed by a CNN and a PointNet architecture, respectively. The resulting outputs are then combined by a novel fusion network, which predicts multiple 3D box hypotheses and their confidences, using the input 3D points as spatial anchors. We evaluate PointFusion on two distinctive datasets: the KITTI dataset that features driving scenes captured with a lidar-camera setup, and the SUN-RGBD dataset that captures indoor environments with RGB-D cameras. Our model is the first one that is able to perform on par or better than the state-of-the-art on these diverse datasets without any dataset-specific mo

del tuning.
****

Scalable Dense Non-Rigid Structure-From-Motion: A Grassmannian Perspective

Suryansh Kumar, Anoop Cherian, Yuchao Dai, Hongdong Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 254-263

This paper addresses the task of dense non-rigid structure-from-motion (NRSfM) using multiple images. State-of-the-art methods to this problem are often hurdled by scalability, expensive computations, and noisy measurements. Further, recent methods to NRSfM usually either assume a small number of sparse feature points or ignore local non-linearities of shape deformations, and thus cannot reliably model complex non-rigid deformations. To address these issues, in this paper, we propose a new approach for dense NRSfM by modeling the problem on a Grassmann manifold. Specifically, we assume the complex non-rigid deformations lie on a union of local linear subspaces both spatially and temporally. This naturally allows for a compact representation of the complex non-rigid deformation over frames. We provide experimental results on several synthetic and real benchmark datasets. The procured results clearly demonstrate that our method, apart from being scalable and more accurate than state-of-the-art methods, is also more robust to noise and generalizes to highly non-linear deformations.
****

GVCNN: Group-View Convolutional Neural Networks for 3D Shape Recognition

Yifan Feng, Zizhao Zhang, Xibin Zhao, Rongrong Ji, Yue Gao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 264-272

3D shape recognition has attracted much attention recently. Its recent advances advocate the usage of deep features and achieve the state-of-the-art performance. However, existing deep features for 3D shape recognition are restricted to a view-to-shape setting, which learns the shape descriptor from the view-level feature directly. Despite the exciting progress on view-based 3D shape description, the intrinsic hierarchical correlation and discriminability among views have not been well exploited, which is important for 3D shape representation. To tackle this issue, in this paper, we propose a group-view convolutional neural network (GVCNN) framework for hierarchical correlation modeling towards discriminative 3D shape description. The proposed GVCNN framework is composed of a hierarchical view-group-shape architecture, i.e., from the view level, the group level and the shape level, which are organized using a grouping strategy. Concretely, we first use an expanded CNN to extract a view level descriptor. Then, a grouping module is introduced to estimate the content discrimination of each view, based on which all views can be splitted into different groups according to their discriminative level. A group level description can be further generated by pooling from view descriptors. Finally, all group level descriptors are combined into the shape level descriptor according to their discriminative weights. Experimental results and comparison with state-of-the-art methods show that our proposed GVCNN method can achieve a significant performance gain on both the 3D shape classification and retrieval tasks.
****

Depth and Transient Imaging With Compressive SPAD Array Cameras

Qilin Sun, Xiong Dun, Yifan Peng, Wolfgang Heidrich; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 273-282

Time-of-flight depth imaging and transient imaging are two imaging modalities that have recently received a lot of interest. Despite much research, existing hardware systems are limited either in terms of temporal resolution or are prohibitively expensive. Arrays of Single Photon Avalanche Diodes (SPADs) promise to fill this gap by providing higher temporal resolution at an affordable cost. Unfortunately SPAD arrays are to date only available in relatively small resolutions. In this work we aim to overcome the spatial resolution limit of SPAD arrays by employing a compressive sensing camera design. Using a DMD and custom optics, we achieve an image resolution of up to 800*400 on SPAD Arrays of resolution 64*32. Using our new data fitting model for the time histograms, we suppress the noise while abstracting the phase and amplitude information, so as to realize a tempo ral resolution of a few tens of picoseconds.

****

GeoNet: Geometric Neural Network for Joint Depth and Surface Normal Estimation

Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasun, Jiaya Jia; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 283-291

In this paper, we propose Geometric Neural Network (GeoNet) to jointly predict depth and surface normal maps from a single image. Building on top of two-stream CNNs, our GeoNet incorporates geometric relation between depth and surface normal via the new depth-to-normal and normal- to-depth networks. Depth-to-normal network exploits the least square solution of surface normal from depth and im- proves its quality with a residual module. Normal-to-depth network, contrarily, refines the depth map based on the con- straints from the surface normal through a kernel regression module, which has no parameter to learn. These two net- works enforce the underlying model to efficiently predict depth and surface normal for high consistency and corre- sponding accuracy. Our experiments on NYU v2 dataset verify that our GeoNet is able to predict geometrically con- sistent depth and normal maps. It achieves top performance on surface normal estimation and is on par with state-of-the- art depth estimation methods.

****

Real-Time Seamless Single Shot 6D Object Pose Prediction

Bugra Tekin, Sudipta N. Sinha, Pascal Fua; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 292-301

We propose a single-shot approach for simultaneously detecting an object in an RGB image and predicting its 6D pose without requiring multiple stages or having to examine multiple hypotheses. Unlike a recently proposed single-shot technique for this task [Kehl et al. 2017] that only predicts an approximate 6D pose that must then be refined, ours is accurate enough not to require additional post-processing. As a result, it is much faster - 50 fps on a Titan X (Pascal) GPU - and more suitable for real-time processing. The key component of our method is a new CNN architecture inspired by [Redmon et al. 2016, Redmon and Farhadi 2017] that directly predicts the 2D image locations of the projected vertices of the object's 3D bounding box. The object's 6D pose is then estimated using a PnP algorithm.   For single object and multiple object pose estimation on the LineMod and Occlusion datasets, our approach substantially outperforms other recent CNN-based approaches [Kehl et al. 2017, Rad and Lepetit 2017] when they are all used without post-processing. During post-processing, a pose refinement step can be used to boost the accuracy of these two methods, but at 10 fps or less, they are much slower than our method.

****

Factoring Shape, Pose, and Layout From the 2D Image of a 3D Scene

Shubham Tulsiani, Saurabh Gupta, David F. Fouhey, Alexei A. Efros, Jitendra Malik; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 302-310

The goal of this paper is to take a single 2D image of a scene and recover the 3D structure in terms of a small set of factors: a layout representing the enclosing surfaces as well as a set of objects represented in terms of shape and pose. We propose a convolutional neural network-based approach to predict this representation and benchmark it on a large dataset of indoor scenes. Our experiments evaluate a number of practical design questions, demonstrate that we can infer this representation, and quantitatively and qualitatively demonstrate its merits compared to alternate representations.

****

Monocular Relative Depth Perception With Web Stereo Data Supervision

Ke Xian, Chunhua Shen, Zhiguo Cao, Hao Lu, Yang Xiao, Ruibo Li, Zhenbo Luo; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 311-320

In this paper we study the problem of monocular relative depth perception in the wild. We introduce a simple yet effective method to automatically generate dense relative depth annotations from web stereo images, and propose a new dataset t hat consists of diverse images as well as corresponding dense relative depth maps. Further, an improved ranking loss is introduced to deal with imbalanced ordinal relations, enforcing the network to focus on a set of hard pairs. Experimental results demonstrate that our proposed approach not only achieves state-of-the-art accuracy of relative depth perception in the wild, but also benefits other dense per-pixel prediction tasks, e.g., metric depth estimation and semantic segmentation.

****

Spline Error Weighting for Robust Visual-Inertial Fusion

Hannes Ovrén, Per-Erik Forssén; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 321-329

In this paper we derive and test a probability-based weighting that can balance residuals of different types in spline fitting. In contrast to previous formulations, the proposed spline error weighting scheme also incorporates a prediction of the approximation error of the spline fit. We demonstrate the effectiveness of the prediction in a synthetic experiment, and apply it to visual-inertial fusion on rolling shutter cameras. This results in a method that can estimate 3D structure with metric scale on generic first-person videos. We also propose a quality measure for spline fitting, that can be used to automatically select the knot spacing. Experiments verify that the obtained trajectory quality corresponds well with the requested quality. Finally, by linearly scaling the weights, we show that the proposed spline error weighting minimizes the estimation errors on real sequences, in terms of scale and end-point errors.

****

Single-Image Depth Estimation Based on Fourier Domain Analysis

Jae-Han Lee, Minhyeok Heo, Kyung-Rae Kim, Chang-Su Kim; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 330-339

We propose a deep learning algorithm for single-image depth estimation based on the Fourier frequency domain analysis. First, we develop a convolutional neural network structure and propose a new loss function, called depth-balanced Euclidean loss, to train the network reliably for a wide range of depths. Then, we generate multiple depth map candidates by cropping input images with various cropping ratios. In general, a cropped image with a small ratio yields depth details more faithfully, while that with a large ratio provides the overall depth distribution more reliably. To take advantage of these complementary properties, we combine the multiple candidates in the frequency domain. Experimental results demonstrate that proposed algorithm provides the state-of-art performance. Furthermore, through the frequency domain analysis, we validate the efficacy of the proposed algorithm in most frequency bands.

****

Unsupervised Learning of Monocular Depth Estimation and Visual Odometry With Deep Feature Reconstruction

Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, Ian Reid; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 340-349

Despite learning based methods showing promising results in single view depth estimation and visual odometry, most existing approaches treat the tasks in a supervised manner. Recent approaches to single view depth estimation explore the possibility of learning without full supervision via minimizing photometric error. In this paper, we explore the use of stereo sequences for learning depth and visual odometry. The use of stereo sequences enables the use of both spatial (between left-right pairs) and temporal (forward backward) photometric warp error, and constrains the scene depth and camera motion to be in a common, real-world scale. At test time our framework is able to estimate single view depth and two-view odometry from a monocular sequence. We also show how we can improve on a standard photometric warp loss by considering a warp of deep features. We show through extensive experiments that: (i) jointly training for single view depth and visual odometry improves depth prediction because of the additional constraint imposed on depths and achieves competitive results for visual odometry; (ii) deep feature-based warping loss improves upon simple photometric warp loss for both sing le view depth estimation and visual odometry. Our method outperforms existing learning based methods on the KITTI driving dataset in both tasks. The source code is available at https://github.com/Huangying-Zhan/Depth-VO-Feat.

****

Detect-and-Track: Efficient Pose Estimation in Videos

Rohit Girdhar, Georgia Gkioxari, Lorenzo Torresani, Manohar Paluri, Du Tran; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 350-359

This paper addresses the problem of estimating and tracking human body keypoints in complex, multi-person video. We propose an extremely lightweight yet highly effective approach that builds upon the latest advancements in human detection and video understanding. Our method operates in two-stages: keypoint estimation in frames or short clips, followed by lightweight tracking to generate keypoint predictions linked over the entire video. For frame-level pose estimation we experiment with Mask R-CNN, as well as our own proposed 3D extension of this model, which leverages temporal information over small clips to generate more robust frame predictions. We conduct extensive ablative experiments on the newly released multi-person video pose estimation benchmark, PoseTrack, to validate various design choices of our model. Our approach achieves an accuracy of 55.2% on the validation and 51.8% on the test set using the Multi-Object Tracking Accuracy (MOTA) metric, and achieves state of the art performance on the ICCV 2017 PoseTrack keypoint tracking challenge.

****

Supervision-by-Registration: An Unsupervised Approach to Improve the Precision of Facial Landmark Detectors

Xuanyi Dong, Shoou-I Yu, Xinshuo Weng, Shih-En Wei, Yi Yang, Yaser Sheikh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 360-368

In this paper, we present supervision-by-registration, an unsupervised approach to improve the precision of facial landmark detectors on both images and video. Our key observation is that the detections of the same landmark in adjacent frames should be coherent with registration, i.e., optical flow. Interestingly, coherency of optical flow is a source of supervision that does not require manual labeling, and can be leveraged during detector training. For example, we can enforce in the training loss function that a detected landmark at frame t-1 followed by optical flow tracking from frame t-1 to frame t should coincide with the location of the detection at frame t. Essentially, supervision-by-registration augments the training loss function with a registration loss, thus training the detector to have output that is not only close to the annotations in labeled images, but also consistent with registration on large amounts of unlabeled videos. End-to-end training with the registration loss is made possible by a differentiable Lucas-Kanade operation, which computes optical flow registration in the forward pass, and back-propagates gradients that encourage temporal coherency in the detector. The output of our method is a more precise image-based facial landmark detector, which can be applied to single images or video. With supervision-by-registration, we demonstrate (1) improvements in facial landmark detection on both images (300W, ALFW) and video (300VW, Youtube-Celebrities), and (2) significant reduction of jittering in video detections.

****

Diversity Regularized Spatiotemporal Attention for Video-Based Person Re-Identification

Shuang Li, Slawomir Bak, Peter Carr, Xiaogang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 369-378

Video-based person re-identification matches video clips of people across non-overlapping cameras. Most existing methods tackle this problem by encoding each video frame in its entirety and computing an aggregate representation across all frames. In practice, people are often partially occluded, which can corrupt the extracted features. Instead, we propose a new spatiotemporal attention model that automatically discovers a diverse set of distinctive body parts. This allows useful information to be extracted from all frames without succumbing to occlusio ns and misalignments. The network learns multiple spatial attention models and employs a diversity regularization term to ensure multiple models do not discover the same body part. Features extracted from local image regions are organized by spatial attention model and are combined using temporal attention. As a result, the network learns latent representations of the face, torso and other body parts using the best available image patches from the entire video sequence. Extensive evaluations on three datasets show that our framework outperforms the state-of-the-art approaches by large margins on multiple metrics.

****

Style Aggregated Network for Facial Landmark Detection

Xuanyi Dong, Yan Yan, Wanli Ouyang, Yi Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 379-388

Recent advances in facial landmark detection achieve success by learning discriminative features from rich deformation of face shapes and poses. Besides the variance of faces themselves, the intrinsic variance of image styles, e.g., grayscale vs. color images, light vs. dark, intense vs. dull, and so on, has constantly been overlooked. This issue becomes inevitable as increasing web images are collected from various sources for training neural networks. In this work, we propose a style-aggregated approach to deal with the large intrinsic variance of image styles for facial landmark detection. Our method transforms original face images to style-aggregated images by a generative adversarial module. The proposed scheme uses the style-aggregated image to maintain face images that are more robust to environmental changes. Then the original face images accompanying with style-aggregated ones play a duet to train a landmark detector which is complementary to each other. In this way, for each face, our method takes two images as input, i.e., one in its original style and the other in the aggregated style. In experiments, we observe that the large variance of image styles would degenerate the performance of facial landmark detectors. Moreover, we show the robustness of our method to the large variance of image styles by comparing to a variant of our approach, in which the generative adversarial module is removed, and no style-aggregated images are used. Our approach is demonstrated to perform well when compared with state-of-the-art algorithms on benchmark datasets AFLW and 300-W. Code is publicly available on GitHub: https://github.com/D-X-Y/SAN

****

Learning Deep Models for Face Anti-Spoofing: Binary or Auxiliary Supervision

Yaojie Liu, Amin Jourabloo, Xiaoming Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 389-398

Face anti-spoofing is crucial to prevent face recognition systems from a security breach. Previous deep learning approaches formulate face anti-spoofing as a binary classification problem. Many of them struggle to grasp adequate spoofing cues and generalize poorly. In this paper, we argue the importance of auxiliary supervision to guide the learning toward discriminative and generalizable cues. A CNN-RNN model is learned to estimate the face depth with pixel-wise supervision, and to estimate rPPG signals with sequence-wise supervision. The estimated depth and rPPG are fused to distinguish live vs. spoof faces. Further, we introduce a new face anti-spoofing database that covers a large range of illumination, subject, and pose variations. Experiments show that our model achieves the state-of-the-art results on both intra- and cross-database testing.

****

Deep Cost-Sensitive and Order-Preserving Feature Learning for Cross-Population Age Estimation

Kai Li, Junliang Xing, Chi Su, Weiming Hu, Yundong Zhang, Stephen Maybank; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 399-408

Facial age estimation from a face image is an important yet very challenging task in computer vision, since humans with different races and/or genders, exhibit quite different patterns in their facial aging processes. To deal with the influence of race and gender, previous methods perform age estimation within each population separately. In practice, however, it is often very difficult to collect and label sufficient data for each population. Therefore, it would be helpful to exploit an existing large labeled dataset of one (source) population to improve the age estimation performance on another (target) population with only a small labeled dataset available. In this work, we propose a Deep Cross-Population (DCP) age estimation model to achieve this goal. In particular, our DCP model develops a two-stage training strategy. First, a novel cost-sensitive multi-task loss function is designed to learn transferable aging features by training on the source population. Second, a novel order-preserving pair-wise loss function is designed to align the aging features of the two populations. By doing so, our DCP model can transfer the knowledge encoded in the source population to the target population. Extensive experiments on the two of the largest benchmark datasets show that our DCP model outperforms several strong baseline methods and many state-of-the-art methods.

****

First-Person Hand Action Benchmark With RGB-D Videos and 3D Hand Pose Annotations

Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, Tae-Kyun Kim; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 409-419

In this work we study the use of 3D hand poses to recognize first-person dynamic hand actions interacting with 3D objects. Towards this goal, we collected RGB-D video sequences comprised of more than 100K frames of 45 daily hand action categories, involving 26 different objects in several hand configurations. To obtain hand pose annotations, we used our own mo-cap system that automatically infers the 3D location of each of the 21 joints of a hand model via 6 magnetic sensors and inverse kinematics. Additionally, we recorded the 6D object poses and provide 3D object models for a subset of hand-object interaction sequences. To the best of our knowledge, this is the first benchmark that enables the study of first-person hand actions with the use of 3D hand poses. We present an extensive experimental evaluation of RGB-D and pose-based action recognition by 18 baselines/state-of-the-art approaches. The impact of using appearance features, poses, and their combinations are measured, and the different training/testing protocols are evaluated. Finally, we assess how ready the 3D hand pose estimation field is when hands are severely occluded by objects in egocentric views and its influence on action recognition. From the results, we see clear benefits of using hand pose as a cue for action recognition compared to other data modalities. Our dataset and experiments can be of interest to communities of 3D hand pose estimation, 6D object pose, and robotics as well as action recognition.

****

A Pose-Sensitive Embedding for Person Re-Identification With Expanded Cross Neighborhood Re-Ranking

M. Saquib Sarfraz, Arne Schumann, Andreas Eberle, Rainer Stiefelhagen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 420-429

Person re-identification is a challenging retrieval task that requires matching a person’s acquired image across non-overlapping camera views. In this paper we propose an effective approach that incorporates both the fine and coarse pose information of the person to learn a discrim- inative embedding. In contrast to the recent direction of explicitly modeling body parts or correcting for misalignment based on these, we show that a rather straightforward inclusion of acquired camera view and/or the detected joint locations into a convolutional neural network helps to learn a very effective representation. To increase retrieval performance, re-ranking techniques based on computed distances have recently gained much attention. We propose a new unsupervised and automatic re-ranking framework that achieves state-of-the-art re-ranking performance. We show that in contrast to the current state-of-the-art re-ranking methods our approach does not require to compute new rank lists for each image pair (e.g., based on reciprocal neighbors) and performs well by using simple direct rank list based comparison or even by just using the already computed euclidean distances between the images. We show that both our learned representation and our re-ranking method achieve state-of-the-art performance on a number of challenging surveillance image and video d

atasets.
****

Disentangling 3D Pose in a Dendritic CNN for Unconstrained 2D Face Alignment

Amit Kumar, Rama Chellappa; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 430-439

Heatmap regression has been used for landmark localization for quite a while now. Most of the methods use a very deep stack of bottleneck modules for heatmap classification stage, followed by heatmap regression to extract the keypoints. In this paper, we present a single dendritic CNN, termed as Pose Conditioned Dendritic Convolution Neural Network (PCD-CNN), where a classification network is followed by a second and modular classification network, trained in an end to end fashion to obtain accurate landmark points. Following a Bayesian formulation, we disentangle the 3D pose of a face image explicitly by conditioning the landmark estimation on pose, making it different from multi-tasking approaches. Extensive experimentation shows that conditioning on pose reduces the localization error by making it agnostic to face pose. The proposed model can be extended to yield variable number of landmark points and hence broadening its applicability to other datasets. Instead of increasing depth or width of the network, we train the CNN efficiently with Mask-Softmax Loss and hard sample mining to achieve upto 15% reduction in error compared to state-of-the-art methods for extreme and medium pose face images from challenging datasets including AFLW, AFW, COFW and IBUG.
****

A Hierarchical Generative Model for Eye Image Synthesis and Eye Gaze Estimation

Kang Wang, Rui Zhao, Qiang Ji; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 440-448

In this work, we introduce a Hierarchical Generative Model (HGM) to enable realistic forward eye image synthe- sis, as well as effective backward eye gaze estimation. The proposed HGM consists of a hierarchical generative shape model (HGSM), and a conditional bidirectional generative adversarial network (c-BiGAN). The HGSM encodes eye ge- ometry knowledge and relates eye gaze with eye shape, while c-BiGAN leverages on big data and captures the dependency between eye shape and eye appearance. As an intermedi- ate component, eye shape connects knowledge-based model (HGSM) with data-driven model (c-BiGAN) and enables bidirectional inference. Through a top-down inference, the HGM can synthesize eye images consistent with the given eye gaze. Through a bottom-up inference, HGM can infer eye gaze effectively from a given eye image. Qualitative and quantitative evaluations on benchmark datasets demonstrate our model’s effectiveness on both eye image synthesis and eye gaze estimation. In addition, the proposed model is not restricted to eye images only. It can be adapted to face images and any shape-appearance related fields.
****

MiCT: Mixed 3D/2D Convolutional Tube for Human Action Recognition

Yizhou Zhou, Xiaoyan Sun, Zheng-Jun Zha, Wenjun Zeng; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 449-458

Human actions in videos are three-dimensional (3D) signals. Recent attempts use 3D convolutional neural networks (CNNs) to explore spatio-temporal information for human action recognition. Though promising, 3D CNNs have not achieved high performanceon on this task with respect to their well-established two-dimensional (2D) counterparts for visual recognition in still images. We argue that the high training complexity of spatio-temporal fusion and the huge memory cost of 3D convolution hinder current 3D CNNs, which stack 3D convolutions layer by layer, by outputting deeper feature maps that are crucial for high-level tasks. We thus propose a Mixed Convolutional Tube (MiCT) that integrates 2D CNNs with the 3D convolution module to generate deeper and more informative feature maps, while reducing training complexity in each round of spatio-temporal fusion. A new end-to-end trainable deep 3D network, MiCT-Net, is also proposed based on the MiCT to better explore spatio-temporal information in human actions. Evaluations on three well-known benchmark datasets (UCF101, Sport-1M and HMDB-51) show that the proposed MiCT-Net significantly outperforms the original 3D CNNs. Compared with state-of-the-art approaches for action recognition on UCF101 and HMDB51, our MiCT-Net

yields the best performance.
****

Learning to Estimate 3D Human Pose and Shape From a Single Color Image
Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, Kostas Daniilidis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 459-468

This work addresses the problem of estimating the full body 3D human pose and shape from a single color image. This is a task where iterative optimization-based solutions have typically prevailed, while Convolutional Networks (ConvNets) have suffered because of the lack of training data and their low resolution 3D predictions. Our work aims to bridge this gap and proposes an efficient and effective direct prediction method based on ConvNets. Central part to our approach is the incorporation of a parametric statistical body shape model (SMPL) within our end-to-end framework. This allows us to get very detailed 3D mesh results, while requiring estimation only of a small number of parameters, making it friendly for direct network prediction. Interestingly, we demonstrate that these parameters can be predicted reliably only from 2D keypoints and masks. These are typical outputs of generic 2D human analysis ConvNets, allowing us to relax the massive requirement that images with 3D shape ground truth are available for training. Simultaneously, by maintaining differentiability, at training time we generate the 3D mesh from the estimated parameters and optimize explicitly for the surface using a 3D per-vertex loss. Finally, a differentiable renderer is employed to project the 3D mesh to the image, which enables further refinement of the network, by optimizing for the consistency of the projection with 2D annotations (i.e., 2D keypoints or masks). The proposed approach outperforms previous baselines on this task and offers an attractive solution for direct prediction of 3D shape from a single color image.
****

Glimpse Clouds: Human Activity Recognition From Unstructured Feature Points
Fabien Baradel, Christian Wolf, Julien Mille, Graham W. Taylor; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 469-478

We propose a method for human activity recognition from RGB data that does not rely on any pose information during test time, and does not explicitly calculate pose information internally. Instead, a visual attention module learns to predict glimpse sequences in each frame. These glimpses correspond to interest points in the scene that are relevant to the classified activities. No spatial coherence is forced on the glimpse locations, which gives the attention module liberty to explore different points at each frame and better optimize the process of scrutinizing visual information. Tracking and sequentially integrating this kind of unstructured data is a challenge, which we address by sep- arating the set of glimpses from a set of recurrent tracking/recognition workers. These workers receive glimpses, jointly performing subsequent motion tracking and activity prediction. The glimpses are soft-assigned to the workers, optimizing coherence of the assignments in space, time and feature space using an external memory module. No hard decisions are taken, i.e. each glimpse point is assigned to all existing workers, albeit with different importance. Our methods outperform the state-of-the-art on the largest human activity recognition dataset available to-date, NTU RGB+D, and on the Northwestern-UCLA Multiview Action 3D Dataset.
****

Context-Aware Deep Feature Compression for High-Speed Visual Tracking
Jongwon Choi, Hyung Jin Chang, Tobias Fischer, Sangdoo Yun, Kyuewang Lee, Jiyeoup Jeong, Yiannis Demiris, Jin Young Choi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 479-488

We propose a new context-aware correlation filter based tracking framework to achieve both high computational speed and state-of-the-art performance among real-time trackers. The major contribution to the high computational speed lies in the proposed deep feature compression that is achieved by a context-aware scheme utilizing multiple expert auto-encoders; a context in our framework refers to the coarse category of the tracking target according to appearance patterns. In the pre-training phase, one expert auto-encoder is trained per category. In the tracking phase, the best expert auto-encoder is selected for a given target, and only this auto-encoder is used. To achieve high tracking performance with the compressed feature map, we introduce extrinsic denoising processes and a new orthogonality loss term for pre-training and fine-tuning of the expert auto-encoders. We validate the proposed context-aware framework through a number of experiments, where our method achieves a comparable performance to state-of-the-art trackers which cannot run in real-time, while running at a significantly fast speed of over 100 fps.

****

Correlation Tracking via Joint Discrimination and Reliability Learning

Chong Sun, Dong Wang, Huchuan Lu, Ming-Hsuan Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 489-497

For visual tracking, an ideal filter learned by the correlation filter (CF) method should take both discrimination and reliability information. However, existing attempts usually focus on the former one while pay less attention to reliability learning. This may make the learned filter be dominated by the unexpected salient regions on the feature map, thereby resulting in model degradation. To address this issue, we propose a novel CF-based optimization problem to jointly model the discrimination and reliability information. First, we treat the filter as the element-wise product of a base filter and a reliability term. The base filter is aimed to learn the discrimination information between the target and backgrounds, and the reliability term encourages the final filter to focus on more reliable regions. Second, we introduce a local response consistency regular term to emphasize equal contributions of different regions and avoid the tracker being dominated by unreliable regions. The proposed optimization problem can be solved using the alternating direction method and speeded up in the Fourier domain. We conduct extensive experiments on the OTB-2013, OTB-2015 and VOT-2016 datasets to evaluate the proposed tracker. Experimental results show that our tracker performs favorably against other state-of-the-art trackers.

****

PhaseNet for Video Frame Interpolation

Simone Meyer, Abdelaziz Djelouah, Brian McWilliams, Alexander Sorkine-Hornung, Markus Gross, Christopher Schroers; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 498-507

Most approaches for video frame interpolation require accurate dense correspondences to synthesize an in-between frame. Therefore, they do not perform well in challenging scenarios with e.g. lighting changes or motion blur. Recent deep learning approaches that rely on kernels to represent motion can only alleviate these problems to some extent. In those cases, methods that use a per-pixel phase-based motion representation have been shown to work well. However, they are only applicable for a limited amount of motion. We propose a new approach, PhaseNet, that is designed to robustly handle challenging scenarios while also coping with larger motion. Our approach consists of a neural network decoder that directly estimates the phase decomposition of the intermediate frame. We show that this is superior to the hand-crafted heuristics previously used in phase-based methods and also compares favorably to recent deep learning based approaches for video frame interpolation on challenging datasets.

****

The Best of Both Worlds: Combining CNNs and Geometric Constraints for Hierarchical Motion Segmentation

Pia Bideau, Aruni RoyChowdhury, Rakesh R. Menon, Erik Learned-Miller; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 508-517

Traditional methods of motion segmentation use powerful geometric constraints to understand motion, but fail to leverage the semantics of high-level image understanding. Modern CNN methods of motion analysis, on the other hand, excel at identifying well-known structures, but may not precisely characterize well-known geometric constraints. In this work, we build a new statistical model of rigid motion flow based on classical perspective projection constraints. We then combine

piecewise rigid motions into complex deformable and articulated objects, guided by semantic segmentation from CNNs and a second ``object-level" statistical model. This combination of classical geometric knowledge combined with the pattern recognition abilities of CNNs yields excellent performance on a wide range of motion segmentation benchmarks, from complex geometric scenes to camouflaged animals.
****

Hyperparameter Optimization for Tracking With Continuous Deep Q-Learning
Xingping Dong, Jianbing Shen, Wenguan Wang, Yu Liu, Ling Shao, Fatih Porikli; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 518-527
Hyperparameters are numerical presets whose values are assigned prior to the commencement of the learning process. Selecting appropriate hyperparameters is critical for the accuracy of tracking algorithms, yet it is difficult to determine their optimal values, in particular, adaptive ones for each specific video sequence. Most hyperparameter optimization algorithms depend on searching a generic range and they are imposed blindly on all sequences. Here, we propose a novel hyperparameter optimization method that can find optimal hyperparameters for a given sequence using an action-prediction network leveraged on Continuous Deep Q-Learning. Since the common state-spaces for object tracking tasks are significantly more complex than the ones in traditional control problems, existing Continuous Deep Q-Learning algorithms cannot be directly applied. To overcome this challenge, we introduce an efficient heuristic to accelerate the convergence behavior. We evaluate our method on several tracking benchmarks and demonstrate its superior performance.
****

Scale-Transferrable Object Detection
Peng Zhou, Bingbing Ni, Cong Geng, Jianguo Hu, Yi Xu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 528-537
Scale problem lies in the heart of object detection. In this work, we develop a novel Scale-Transferrable Detection Network (STDN) for detecting multi-scale objects in images. In contrast to previous methods that simply combine object predictions from multiple feature maps from different network depths, the proposed network is equipped with embedded super-resolution layers (named as scale-transfer layer/module in this work) to explicitly explore the inter-scale consistency nature across multiple detection scales. Scale-transfer module naturally fits the base network with little computational cost. This module is further integrated with a dense convolutional network (DenseNet) to yield a one-stage object detector. We evaluate our proposed architecture on PASCAL VOC 2007 and MS COCO benchmark tasks and STDN obtains significant improvements over the comparable state-of-the-art detection models.
****

A Prior-Less Method for Multi-Face Tracking in Unconstrained Videos
Chung-Ching Lin, Ying Hung; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 538-547
This paper presents a prior-less method for tracking and clustering an unknown number of human faces and maintaining their individual identities in unconstrained videos. The key challenge is to accurately track faces with partial occlusion and drastic appearance changes in multiple shots resulting from significant variations of makeup, facial expression, head pose and illumination. To address this challenge, we propose a new multi-face tracking and re-identification algorithm, which provides high accuracy in face association in the entire video with automatic cluster number generation, and is robust to outliers. We develop a co-occurrence model of multiple body parts to seamlessly create face tracklets, and recursively link tracklets to construct a graph for extracting clusters. A Gaussian Process model is introduced to compensate the deep feature insufficiency, and is further used to refine the linking results. The advantages of the proposed algorithm are demonstrated using a variety of challenging music videos and newly introduced body-worn camera videos. The proposed method obtains significant improvements over the state of the art [51], while relying less on handling video-spec

ific prior information to achieve high performance.

****

End-to-End Flow Correlation Tracking With Spatial-Temporal Attention

Zheng Zhu, Wei Wu, Wei Zou, Junjie Yan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 548-557

Discriminative correlation filters (DCF) with deep convolutional features have achieved favorable performance in recent tracking benchmarks. However, most of existing DCF trackers only consider appearance features of current frame, and hardly benefit from motion and inter-frame information. The lack of temporal information degrades the tracking performance during challenges such as partial occlusion and deformation. In this paper, we propose the FlowTrack, which focuses on making use of the rich flow information in consecutive frames to improve the feature representation and the tracking accuracy. The FlowTrack formulates individual components, including optical flow estimation, feature extraction, aggregation and correlation filters tracking as special layers in network. To the best of our knowledge, this is the first work to jointly train flow and tracking task in deep learning framework. Then the historical feature maps at predefined intervals are warped and aggregated with current ones by the guiding of flow. For adaptive aggregation, we propose a novel spatial-temporal attention mechanism. In experiments, the proposed method achieves leading performance on OTB2013, OTB2015, VOT2015 and VOT2016.

****

Deep Texture Manifold for Ground Terrain Recognition

Jia Xue, Hang Zhang, Kristin Dana; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 558-567

We present a texture network called Deep Encoding Pooling Network (DEP) for the task of ground terrain recognition. Recognition of ground terrain is an important task in establishing robot or vehicular control parameters, as well as for localization within an outdoor environment. The architecture of DEP integrates orderless texture details and local spatial information and the performance of DEP surpasses state-of-the-art methods for this task. The GTOS database (comprised of over 30,000 images of 40 classes of ground terrain in outdoor scenes) enables supervised recognition. For evaluation under realistic conditions, we use test images that are not from the existing GTOS dataset, but are instead from hand-held mobile phone videos of similar terrain. This new evaluation dataset, GTOS-mobile, consists of 81 videos of 31 classes of ground terrain such as grass, gravel, asphalt and sand. The resultant network shows excellent performance not only for GTOS-mobile, but also for more general databases (MINC and DTD). Leveraging the discriminant features learned from this network, we build a new texture manifold called DEP-manifold. We learn a parametric distribution in feature space in a fully supervised manner, which gives the distance relationship among classes and provides a means to implicitly represent ambiguous class boundaries. The source code and database are publicly available.

****

Learning Superpixels With Segmentation-Aware Affinity Loss

Wei-Chih Tu, Ming-Yu Liu, Varun Jampani, Deqing Sun, Shao-Yi Chien, Ming-Hsuan Yang, Jan Kautz; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 568-576

Superpixel segmentation has been widely used in many computer vision tasks. Existing superpixel algorithms are mainly based on hand-crafted features, which often fail to preserve weak object boundaries. In this work, we leverage deep neural networks to facilitate extracting superpixels from images. We show a simple integration of deep features with existing superpixel algorithms does not result in better performance as these features do not model segmentation. Instead, we propose a segmentation-aware affinity learning approach for superpixel segmentation. Specifically, we propose a new loss function that takes the segmentation error into account for affinity learning. We also develop the Pixel Affinity Net for affinity prediction. Extensive experimental results show that the proposed algorithm based on the learned segmentation-aware loss performs favorably against the state-of-the-art methods. We also demonstrate the use of the learned superpixel s in numerous vision applications with consistent improvements.

****

Interactive Image Segmentation With Latent Diversity

Zhuwen Li, Qifeng Chen, Vladlen Koltun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 577-585

Interactive image segmentation is characterized by multimodality. When the user clicks on a door, do they intend to select the door or the whole house? We present an end-to-end learning approach to interactive image segmentation that tackles this ambiguity. Our architecture couples two convolutional networks. The first is trained to synthesize a diverse set of plausible segmentations that conform to the user's input. The second is trained to select among these. By selecting a single solution, our approach retains compatibility with existing interactive segmentation interfaces. By synthesizing multiple diverse solutions before selecting one, the architecture is given the representational power to explore the multimodal solution space. We show that the proposed approach outperforms existing methods for interactive image segmentation, including prior work that applied convolutional networks to this problem, while being much faster.

****

The Unreasonable Effectiveness of Deep Features as a Perceptual Metric

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, Oliver Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 586-595

While it is nearly effortless for humans to quickly assess the perceptual similarity between two images, the underlying processes are thought to be quite complex. Despite this, the most widely used perceptual metrics today, such as PSNR and SSIM, are simple, shallow functions, and fail to account for many nuances of human perception. Recently, the deep learning community has found that features of the VGG network trained on ImageNet classification has been remarkably useful as a training loss for image synthesis. But how perceptual are these so-called ``perceptual losses"? What elements are critical for their success? To answer these questions, we introduce a new dataset of human perceptual similarity judgments. We systematically evaluate deep features across different architectures and tasks and compare them with classic metrics. We find that deep features outperform all previous metrics by large margins on our dataset. More surprisingly, this result is not restricted to ImageNet-trained VGG features, but holds across different deep architectures and levels of supervision (supervised, self-supervised, or even unsupervised). Our results suggest that perceptual similarity is an emergent property shared across deep visual representations.

****

Local Descriptors Optimized for Average Precision

Kun He, Yan Lu, Stan Sclaroff; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 596-605

Extraction of local feature descriptors is a vital stage in the solution pipelines for numerous computer vision tasks. Learning-based approaches improve performance in certain tasks, but still cannot replace handcrafted features in general. In this paper, we improve the learning of local feature descriptors by optimizing the performance of descriptor matching, which is a common stage that follows descriptor extraction in local feature based pipelines, and can be formulated as nearest neighbor retrieval. Specifically, we directly optimize a ranking-based retrieval performance metric, Average Precision, using deep neural networks. This general-purpose solution can also be viewed as a listwise learning to rank approach, which is advantageous compared to recent local ranking approaches. On standard benchmarks, descriptors learned with our formulation achieve state-of-the-art results in patch verification, patch retrieval, and image matching.

****

Recovering Realistic Texture in Image Super-Resolution by Deep Spatial Feature Transform

Xintao Wang, Ke Yu, Chao Dong, Chen Change Loy; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 606-615

Despite that convolutional neural networks (CNN) have recently demonstrated high

-quality reconstruction for single-image super-resolution (SR), recovering natural and realistic texture remains a challenging problem. In this paper, we show that it is possible to recover textures faithful to semantic classes. In particular, we only need to modulate features of a few intermediate layers in a single network conditioned on semantic segmentation probability maps. This is made possible through a novel Spatial Feature Transform (SFT) layer that generates affine transformation parameters for spatial-wise feature modulation. SFT layers can be trained end-to-end together with the SR network using the same loss function. During testing, it accepts an input image of arbitrary size and generates a high-resolution image with just a single forward pass conditioned on the categorical priors. Our final results show that an SR network equipped with SFT can generate more realistic and visually pleasing textures in comparison to state-of-the-art SRGAN and EnhanceNet.

****

Deep Extreme Cut: From Extreme Points to Object Segmentation

Kevis-Kokitsi Maninis, Sergi Caelles, Jordi Pont-Tuset, Luc Van Gool; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 616-625

This paper explores the use of extreme points in an object (left-most, right-most, top, bottom pixels) as input to obtain precise object segmentation for images and videos. We do so by adding an extra channel to the image in the input of a convolutional neural network (CNN), which contains a Gaussian centered in each of the extreme points. The CNN learns to transform this information into a segmentation of an object that matches those extreme points. We demonstrate the usefulness of this approach for guided segmentation (grabcut-style), interactive segmentation, video object segmentation, and dense segmentation annotation. We show that we obtain the most precise results to date, also with less user input, in an extensive and varied selection of benchmarks and datasets. All our models and code are publicly available on http://www.vision.ee.ethz.ch/~cvlsegmentation/dextr.

****

Learning to Parse Wireframes in Images of Man-Made Environments

Kun Huang, Yifan Wang, Zihan Zhou, Tianjiao Ding, Shenghua Gao, Yi Ma; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 626-635

In this paper, we propose a learning-based approach to the task of automatically extracting a "wireframe" representation for images of cluttered man-made environments. The wireframe contains all salient straight lines and their junctions of the scene that encode efficiently and accurately large-scale geometry and object shapes. To this end, we have built a very large new dataset of over 5,000 images with wireframes thoroughly labelled by humans. We have proposed two convolutional neural networks that are suitable for extracting junctions and lines with large spatial support, respectively. The networks trained on our dataset have achieved significantly better performance than state-of-the-art methods for junction detection and line segment detection, respectively. We have conducted extensive experiments to evaluate quantitatively and qualitatively the wireframes obtained by our method, and have convincingly shown that effectively and efficiently parsing wireframes for images of man-made environments is a feasible goal within reach. Such wireframes could benefit many important visual tasks such as feature correspondence, 3D reconstruction, vision-based mapping, localization, and navigation.

****

Occlusion-Aware Rolling Shutter Rectification of 3D Scenes

Subeesh Vasu, Mahesh Mohan M. R., A. N. Rajagopalan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 636-645

A vast majority of contemporary cameras employ rolling shutter (RS) mechanism to capture images. Due to the sequential mechanism, images acquired with a moving camera are subjected to rolling shutter effect which manifests as geometric distortions. In this work, we consider the specific scenario of a fast moving camera wherein the rolling shutter distortions not only are predominant but also becom e depth-dependent which in turn results in intra-frame occlusions. To this end, we develop a first-of-its-kind pipeline to recover the latent image of a 3D scene from a set of such RS distorted images. The proposed approach sequentially recovers both the camera motion and scene structure while accounting for RS and occlusion effects. Subsequently, we perform depth and occlusion-aware rectification of RS images to yield the desired latent image. Our experiments on synthetic and real image sequences reveal that the proposed approach achieves state-of-the-art results.

****

Content-Sensitive Supervoxels via Uniform Tessellations on Video Manifolds
Ran Yi, Yong-Jin Liu, Yu-Kun Lai; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 646-655
Supervoxels are perceptually meaningful atomic regions in videos, obtained by grouping voxels that exhibit coherence in both appearance and motion. In this paper, we propose content-sensitive supervoxels (CSS), which are regularly-shaped 3D primitive volumes that possess the following characteristic: they are typically larger and longer in content-sparse regions (i.e., with homogeneous appearance and motion), and smaller and shorter in content-dense regions (i.e., with high variation of appearance and/or motion). To compute CSS, we map a video X to a 3-dimensional manifold M embedded in R^6, whose volume elements give a good measure of the content density in X. We propose an efficient Lloyd-like method with a splitting-merging scheme to compute a uniform tessellation on M, which induces the CSS in X. Theoretically our method has a good competitive ratio O(1). We also present a simple extension of CSS to stream CSS for processing long videos that cannot be loaded into main memory at once. We evaluate CSS, stream CSS and seven representative supervoxel methods on four video datasets. The results show that our method outperforms existing supervoxel methods.

****

Intrinsic Image Transformation via Scale Space Decomposition
Lechao Cheng, Chengyi Zhang, Zicheng Liao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 656-665
We introduce a new network structure for decomposing an image into its intrinsic albedo and shading. We treat this as an image-to-image transformation problem and explore the scale space of the input and output. By expanding the output images (albedo and shading) into their Laplacian pyramid components, we develop a multi-channel network structure that learns the image-to-image transformation function in successive frequency bands in parallel, within each channel is a fully convolutional neural network with skip connections. This network structure is general and extensible, and has demonstrated excellent performance on the intrinsic image decomposition problem. We evaluate the network on two benchmark datasets: the MPI-Sintel dataset and the MIT Intrinsic Images dataset. Both quantitative and qualitative results show our model delivers a clear progression over state-of-the-art.

****

Learned Shape-Tailored Descriptors for Segmentation
Naeemullah Khan, Ganesh Sundaramoorthi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 666-674
We address the problem of texture segmentation by grouping dense pixel-wise descriptors. We introduce and construct learned Shape-Tailored Descriptors that aggregate image statistics only within regions of interest to avoid mixing statistics of different textures, and that are invariant to complex nuisances (e.g., illumination, perspective and deformations). This is accomplished by training a neural network to discriminate base shape-tailored descriptors of oriented gradients at various scales. These descriptors are defined through partial differential equations to obtain data at various scales in arbitrarily shaped regions. We formulate and optimize a joint optimization problem in the segmentation and descriptors to discriminate these base descriptors using the learned metric, equivalent to grouping learned descriptors. We test the method on datasets to illustrate the effect of both the shape-tailored and learned properties of the descriptors. Experiments show that the descriptors learned on a small dataset of segmented ima

ges generalize well to unseen textures in other datasets, showing the generic nature of these descriptors. We show stateof- the-art results on texture segmentation benchmarks.

****

PAD-Net: Multi-Tasks Guided Prediction-and-Distillation Network for Simultaneous Depth Estimation and Scene Parsing

Dan Xu, Wanli Ouyang, Xiaogang Wang, Nicu Sebe; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 675-684

Depth estimation and scene parsing are two particularly important tasks in visual scene understanding. In this paper we tackle the problem of simultaneous depth estimation and scene parsing in a joint CNN. The task can be typically treated as a deep multi-task learning problem [42]. Different from previous methods directly optimizing multiple tasks given the input training data, this paper proposes a novel multi-task guided prediction-and-distillation network (PAD-Net), which first predicts a set of intermediate auxiliary tasks ranging from low level to high level, and then the predictions from these intermediate auxiliary tasks are utilized as multi-modal input via our proposed multi-modal distillation modules for the final tasks. During the joint learning, the intermediate tasks not only act as supervision for learning more robust deep representations but also provide rich multi-modal information for improving the final tasks. Extensive experiments are conducted on two challenging datasets (i.e. NYUD-v2 and Cityscapes) for both the depth estimation and scene parsing tasks, demonstrating the effectiveness of the proposed approach.

****

Multi-Image Semantic Matching by Mining Consistent Features

Qianqian Wang, Xiaowei Zhou, Kostas Daniilidis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 685-694

This work proposes a multi-image matching method to estimate semantic correspondences across multiple images. In contrast to the previous methods that optimize all pairwise correspondences, the proposed method identifies and matches only a sparse set of reliable features in the image collection. In this way, the proposed method is able to prune nonrepeatable features and also highly scalable to handle thousands of images. We additionally propose a low-rank constraint to ensure the geometric consistency of feature correspondences over the whole image collection. Besides the competitive performance on multi-graph matching and semantic flow benchmarks, we also demonstrate the applicability of the proposed method for reconstructing object-class models and discovering object-class landmarks from images without using any annotation.

****

Density-Aware Single Image De-Raining Using a Multi-Stream Dense Network

He Zhang, Vishal M. Patel; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 695-704

Single image rain streak removal is an extremely challenging problem due to the presence of non-uniform rain densities in images. We present a novel density-aware multi-stream densely connected convolutional neural network-based algorithm, called DID-MDN, for joint rain density estimation and de-raining. The proposed method enables the network itself to automatically determine the rain-density information and then efficiently remove the corresponding rain-streaks guided by the estimated rain-density label. To better characterize rain-streaks with different scales and shapes, a multi-stream densely connected de-raining network is proposed which efficiently leverages features from different scales. Furthermore, a new dataset containing images with rain-density labels is created and used to train the proposed density-aware network. Extensive experiments on synthetic and real datasets demonstrate that the proposed method achieves significant improvements over the recent state-of-the-art methods. In addition, an ablation study is performed to demonstrate the improvements obtained by different modules in the proposed method.

****

Joint Cuts and Matching of Partitions in One Graph

Tianshu Yu, Junchi Yan, Jieyi Zhao, Baoxin Li; Proceedings of the IEEE Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 705-713

As two fundamental problems, graph cuts and graph matching have been intensively investigated over the decades, resulting in vast literature in these two topics respectively. However the way of jointly applying and solving graph cuts and matching receives few attention. In this paper, we first formalize the problem of simultaneously cutting a graph into two partitions i.e. graph cuts and establishing their correspondence i.e. graph matching. Then we develop an optimization algorithm by updating matching and cutting alternatively, provided with theoretical analysis. The efficacy of our algorithm is verified on both synthetic dataset and real-world images containing similar regions or structures.

****

Progressive Attention Guided Recurrent Network for Salient Object Detection

Xiaoning Zhang, Tiantian Wang, Jinqing Qi, Huchuan Lu, Gang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 714-722

Effective convolutional features play an important role in saliency estimation but how to learn powerful features for saliency is still a challenging task. FCN-based methods directly apply multi-level convolutional features without distinction, which leads to sub-optimal results due to the distraction from redundant details. In this paper, we propose a novel attention guided network which selectively integrates multi-level contextual information in a progressive manner. Attentive features generated by our network can alleviate distraction of background thus achieve better performance. On the other hand, it is observed that most of existing algorithms conduct salient object detection by exploiting side-output features of the backbone feature extraction network. However, shallower layers of backbone network lack the ability to obtain global semantic information, which limits the effective feature learning. To address the problem, we introduce multi-path recurrent feedback to enhance our proposed progressive attention driven framework. Through multi-path recurrent connections, global semantic information from the top convolutional layer is transferred to shallower layers, which intrinsically refines the entire network. Experimental results on six benchmark datasets demonstrate that our algorithm performs favorably against the state-of-the-art approaches.

****

Fast and Accurate Single Image Super-Resolution via Information Distillation Network

Zheng Hui, Xiumei Wang, Xinbo Gao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 723-731

Recently, deep convolutional neural networks (CNNs) have been demonstrated remarkable progress on single image super-resolution. However, as the depth and width of the networks increase, CNN-based super-resolution methods have been faced with the challenges of computational complexity and memory consumption in practice. In order to solve the above questions, we propose a deep but compact convolutional network to directly reconstruct the high resolution image from the original low resolution image. In general, the proposed model consists of three parts, which are feature extraction block, stacked information distillation blocks and reconstruction block respectively. By combining an enhancement unit with a compression unit into a distillation block, the local long and short-path features can be effectively extracted. Specifically, the proposed enhancement unit mixes together two different types of features and the compression unit distills more useful information for the sequential blocks. In addition, the proposed network has the advantage of fast execution due to the comparatively few number of filters per layer and the use of group convolution. Experimental results demonstrate that the proposed method is superior to the state-of-the-art methods, especially in terms of time performance.

****

Hallucinated-IQA: No-Reference Image Quality Assessment via Adversarial Learning

Kwan-Yee Lin, Guanxiang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 732-741

No-reference image quality assessment (NR-IQA) is a fundamental yet challenging task in low-level computer vision community. The difficulty is particularly pronounced for the limited information, for which the corresponding reference for comparison is typically absent. Although various feature extraction mechanisms have been leveraged from natural scene statistics to deep neural networks in previous methods, the performance bottleneck still exists. In this work, we propose a hallucination-guided quality regression network to address the issue. We firstly generate a hallucinated reference constrained on the distorted image, to compensate the absence of the true reference. Then, we pair the information of hallucinated reference with the distorted image, and forward them to the regressor to learn the perceptual discrepancy with the guidance of an implicit ranking relationship within the generator, and therefore produce the precise quality prediction. To demonstrate the effectiveness of our approach, comprehensive experiments are evaluated on four popular image quality assessment benchmarks. Our method significantly outperforms all the previous state-of-the-art methods by large margins. The code and model are publicly available on the project page https://kwanyeelin.github.io/projects/HIQA/HIQA.html

****

NAG: Network for Adversary Generation

Konda Reddy Mopuri, Utkarsh Ojha, Utsav Garg, R. Venkatesh Babu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 742-751

Adversarial perturbations can pose a serious threat for deploying machine learning systems. Recent works have shown existence of image-agnostic perturbations that can fool classifiers over most natural images. Existing methods present optimization approaches that solve for a fooling objective with an imperceptibility constraint to craft the perturbations. However, for a given classifier, they generate one perturbation at a time, which is a single instance from the manifold of adversarial perturbations. Also, in order to build robust models, it is essential to explore the manifold of adversarial perturbations. In this paper, we propose for the first time, a generative approach to model the distribution of adversarial perturbations. The architecture of the proposed model is inspired from that of GANs and is trained using fooling and diversity objectives. Our trained generator network attempts to capture the distribution of adversarial perturbations for a given classifier and readily generates a wide variety of such perturbations. Our experimental evaluation demonstrates that perturbations crafted by our model (i) achieve state-of-the-art fooling rates, (ii) exhibit wide variety and (iii) deliver excellent cross model generalizability. Our work can be deemed as an important step in the process of inferring about the complex manifolds of adversarial perturbations.

****

Dynamic-Structured Semantic Propagation Network

Xiaodan Liang, Hongfei Zhou, Eric Xing; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 752-761

Semantic concept hierarchy is yet under-explored for semantic segmentation due to the inefficiency and complicated optimization of incorporating structural inference into the dense prediction. This lack of modeling dependencies among concepts severely limits the generalization capability of segmentation models for open set large-scale vocabularies. Prior works thus must tune highly-specified models for each task due to the label discrepancy across datasets. In this paper, we propose a Dynamic-Structured Semantic Propagation Network (DSSPN) that builds a semantic neuron graph to explicitly incorporate the concept hierarchy into dynamic network construction, leading to an interpretable reasoning process. Each neuron for one super-class (eg food) represents the instantiated module for recognizing among fine-grained child concepts (eg editable fruit or pizza), and then its learned features flow into the child neurons (eg distinguishing between orange or apple) for hierarchical categorization in finer levels. A dense semantic-enhanced neural block propagates the learned knowledge of all ancestral neurons into each fine-grained child neuron for progressive feature evolving. During training, DSSPN performs the dynamic-structured neuron computational graph by only activating a sub-graph of neurons for each image. Another merit of such semantic ex

plainable structure is the ability to learn a unified model concurrently on diverse datasets by selectively activating different neuron sub-graphs for each annotation at each step. Extensive experiments on four public semantic segmentation datasets (i.e. ADE20K, COCO-Stuff, Cityscape and Mapillary) demonstrate the superiority of DSSPN, and a universal segmentation model that is jointly trained on diverse datasets can surpass the common fine-tuning scheme for exploiting multi-domain knowledge.
****

Cross-Domain Self-Supervised Multi-Task Feature Learning Using Synthetic Imagery
Zhongzheng Ren, Yong Jae Lee; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 762-771
In human learning, it is common to use multiple sources of information jointly. However, most existing feature learning approaches learn from only a single task. In this paper, we propose a novel multi-task deep network to learn generalizable high-level visual representations. Since multi-task learning requires annotations for multiple properties of the same training instance, we look to synthetic images to train our network. To overcome the domain difference between real and synthetic data, we employ an unsupervised feature space domain adaptation method based on adversarial learning. Given an input synthetic RGB image, our network simultaneously predicts its surface normal, depth, and instance contour, while also minimizing the feature space domain differences between real and synthetic data. Through extensive experiments, we demonstrate that our network learns more transferable representations compared to single-task baselines. Our learned representation produces state-of-the-art transfer learning results on PASCAL VOC 2007 classification and 2012 detection.
****

A Two-Step Disentanglement Method
Naama Hadad, Lior Wolf, Moni Shahar; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 772-780
We address the problem of disentanglement of factors that generate a given data into those that are correlated with the labeling and those that are not. Our solution is simpler than previous solutions and employs adversarial training. First, the part of the data that is correlated with the labels is extracted by training a classifier. Then, the other part is extracted such that it enables the reconstruction of the original data but does not contain label information. The utility of the new method is demonstrated on visual datasets as well as on financial data. Our code is available at https://github.com/naamahadad/A-Two-Step-Disentanglement-Method.
****

Robust Facial Landmark Detection via a Fully-Convolutional Local-Global Context Network
Daniel Merget, Matthias Rock, Gerhard Rigoll; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 781-790
While fully-convolutional neural networks are very strong at modeling local features, they fail to aggregate global context due to their constrained receptive field. Modern methods typically address the lack of global context by introducing cascades, pooling, or by fitting a statistical model. In this work, we propose a new approach that introduces global context into a fully-convolutional neural network directly. The key concept is an implicit kernel convolution within the network. The kernel convolution blurs the output of a local-context subnet, which is then refined by a global-context subnet using dilated convolutions. The kernel convolution is crucial for the convergence of the network because it smoothens the gradients and reduces overfitting. In a postprocessing step, a simple PCA-based 2D shape model is fitted to the network output in order to filter outliers. Our experiments demonstrate the effectiveness of our approach, outperforming several state-of-the-art methods in facial landmark detection.
****

Decorrelated Batch Normalization
Lei Huang, Dawei Yang, Bo Lang, Jia Deng; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 791-800

Batch Normalization (BN) is capable of accelerating the training of deep models by centering and scaling activations within mini-batches. In this work, we propose Decorrelated Batch Normalization (DBN), which not just centers and scales activations but whitens them. We explore multiple whitening techniques, and find that PCA whitening causes a problem we call stochastic axis swapping, which is detrimental to learning. We show that ZCA whitening does not suffer from this problem, permitting successful learning. DBN retains the desirable qualities of BN and further improves BN's optimization efficiency and generalization ability. We design comprehensive experiments to show that DBN can improve the performance of BN on multilayer perceptrons and convolutional neural networks. Furthermore, we consistently improve the accuracy of residual networks on CIFAR-10, CIFAR-100, and ImageNet.

****

Learning to Sketch With Shortcut Cycle Consistency

Jifei Song, Kaiyue Pang, Yi-Zhe Song, Tao Xiang, Timothy M. Hospedales; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 801-810

To see is to sketch -- free-hand sketching naturally builds ties between human and machine vision. In this paper, we present a novel approach for translating an object photo to a sketch, mimicking the human sketching process. This is an extremely challenging task because the photo and sketch domains differ significantly. Furthermore, human sketches exhibit various levels of sophistication and abstraction even when depicting the same object instance in a reference photo. This means that even if photo-sketch pairs are available, they only provide weak supervision signal to learn a translation model. Compared with existing supervised approaches that solve the problem of D(E(photo)) -> sketch, where E(·) and D(·) denote encoder and decoder respectively, we take advantage of the inverse problem (e.g., D(E(sketch)) -> photo), and combine with the unsupervised learning tasks of within-domain reconstruction, all within a multi-task learning framework. Compared with existing unsupervised approaches based on cycle consistency (i.e., D(E(D(E(photo)))) -> photo), we introduce a shortcut consistency enforced at the encoder bottleneck (e.g., D(E(photo)) -> photo) to exploit the additional self-supervision. Both qualitative and quantitative results show that the proposed model is superior to a number of state-of-the-art alternatives. We also show that the synthetic sketches can be used to train a better fine-grained sketch-based image retrieval (FG-SBIR) model, effectively alleviating the problem of sketch data scarcity.

****

Towards a Mathematical Understanding of the Difficulty in Learning With Feedforward Neural Networks

Hao Shen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 811-820

Training deep neural networks for solving machine learning problems is one great challenge in the field, mainly due to its associated optimisation problem being highly non-convex. Recent developments have suggested that many training algorithms do not suffer from undesired local minima under certain scenario, and consequently led to great efforts in pursuing mathematical explanations for such observations. This work provides an alternative mathematical understanding of the challenge from a smooth optimisation perspective. By assuming exact learning of finite samples, sufficient conditions are identified via a critical point analysis to ensure any local minimum to be globally minimal as well. Furthermore, a state of the art algorithm, known as the Generalised Gauss-Newton (GGN) algorithm, is rigorously revisited as an approximate Newton's algorithm, which shares the property of being locally quadratically convergent to a global minimum under the condition of exact learning.

****

FaceID-GAN: Learning a Symmetry Three-Player GAN for Identity-Preserving Face Synthesis

Yujun Shen, Ping Luo, Junjie Yan, Xiaogang Wang, Xiaoou Tang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 82

1-830

Face synthesis has achieved advanced development by using generative adversarial networks (GANs). Existing methods typically formulate GAN as a two-player game, where a discriminator distinguishes face images from the real and synthesized domains, while a generator reduces its discriminativeness by synthesizing a face of photo-realistic quality. Their competition converges when the discriminator is unable to differentiate these two domains. Unlike two-player GANs, this work generates identity-preserving faces by proposing FaceID-GAN, which treats a classifier of face identity as the third player, competing with the generator by distinguishing the identities of the real and synthesized faces (see Fig.1). A stationary point is reached when the generator produces faces that have high quality as well as preserve identity. Instead of simply modeling the identity classifier as an additional discriminator, FaceID-GAN is formulated by satisfying information symmetry, which ensures that the real and synthesized images are projected into the same feature space. In other words, the identity classifier is used to extract identity features from both input (real) and output (synthesized) face images of the generator, substantially alleviating training difficulty of GAN. Extensive experiments show that FaceID-GAN is able to generate faces of arbitrary viewpoint while preserve identity, outperforming recent advanced approaches.

****

A Constrained Deep Neural Network for Ordinal Regression

Yanzhu Liu, Adams Wai Kin Kong, Chi Keong Goh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 831-839

Ordinal regression is a supervised learning problem aiming to classify instances into ordinal categories. It is challenging to automatically extract high-level features for representing intraclass information and interclass ordinal relationship simultaneously. This paper proposes a constrained optimization formulation for the ordinal regression problem which minimizes the negative loglikelihood for multiple categories constrained by the order relationship between instances. Mathematically, it is equivalent to an unconstrained formulation with a pairwise regularizer. An implementation based on the CNN framework is proposed to solve the problem such that high-level features can be extracted automatically, and the optimal solution can be learned through the traditional back-propagation method. The proposed pairwise constraints make the algorithm work even on small datasets, and a proposed efficient implementation make it be scalable for large datasets. Experimental results on four real-world benchmarks demonstrate that the proposed algorithm outperforms the traditional deep learning approaches and other state-of-the-art approaches based on hand-crafted features.

****

Modulated Convolutional Networks

Xiaodi Wang, Baochang Zhang, Ce Li, Rongrong Ji, Jungong Han, Xianbin Cao, Jianzhuang Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 840-848

Despite great effectiveness of very deep and wide Convolutional Neural Networks (CNNs) in various computer vision tasks, the significant cost in terms of storage requirement of such networks impedes the deployment on computationally limited devices. In this paper, we propose new Modulated Convolutional Networks (MCNs) to improve the portability of CNNs via binarized filters. In MCNs, we propose a new loss function which considers the filter loss, center loss and softmax loss in an end-to-end framework. We first introduce modulation filters (M-Filters) to recover the unbinarized filters, which leads to a new architecture to calculate the network model. The convolution operation is further approximated by considering intra-class compactness in the loss function. As a result, our MCNs can reduce the size of required storage space of convolutional filters by a factor of 32, in contrast to the full-precision model, while achieving much better performances than state-of-the-art binarized models. Most importantly, MCNs achieve a comparable performance to the full-precision ResNets and Wide-ResNets. The code will be available publicly soon.

****

Learning Steerable Filters for Rotation Equivariant CNNs

Maurice Weiler, Fred A. Hamprecht, Martin Storath; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 849-858

In many machine learning tasks it is desirable that a model's prediction transforms in an equivariant way under transformations of its input. Convolutional neural networks (CNNs) implement translational equivariance by construction; for other transformations, however, they are compelled to learn the proper mapping. In this work, we develop Steerable Filter CNNs (SFCNNs) which achieve joint equivariance under translations and rotations by design. The proposed architecture employs steerable filters to efficiently compute orientation dependent responses for many orientations without suffering interpolation artifacts from filter rotation. We utilize group convolutions which guarantee an equivariant mapping. In addition, we generalize He's weight initialization scheme to filters which are defined as a linear combination of a system of atomic filters. Numerical experiments show a substantial enhancement of the sample complexity with a growing number of sampled filter orientations and confirm that the network generalizes learned patterns over orientations. The proposed approach achieves state-of-the-art on the rotated MNIST benchmark and on the ISBI 2012 2D EM segmentation challenge.

****

Efficient Interactive Annotation of Segmentation Datasets With Polygon-RNN++

David Acuna, Huan Ling, Amlan Kar, Sanja Fidler; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 859-868

Manually labeling datasets with object masks is extremely time consuming. In this work, we follow the idea of Polygon-RNN to produce polygonal annotations of objects interactively using humans-in-the-loop. We introduce several important improvements to the model: 1) we design a new CNN encoder architecture, 2) show how to effectively train the model with Reinforcement Learning, and 3) significantly increase the output resolution using a Graph Neural Network, allowing the model to accurately annotate high-resolution objects in images. Extensive evaluation on the Cityscapes dataset shows that our model, which we refer to as Polygon-RNN++, significantly outperforms the original model in both automatic (10% absolute and 16% relative improvement in mean IoU) and interactive modes (requiring 50% fewer clicks by annotators). We further analyze the cross-domain scenario in which our model is trained on one dataset, and used out of the box on datasets from varying domains. The results show that Polygon-RNN++ exhibits powerful generalization capabilities, achieving significant improvements over existing pixel-wise methods. Using simple online fine-tuning we further achieve a high reduction in annotation time for new datasets, moving a step closer towards an interactive annotation tool to be used in practice.

****

SplineCNN: Fast Geometric Deep Learning With Continuous B-Spline Kernels

Matthias Fey, Jan Eric Lenssen, Frank Weichert, Heinrich Müller; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 869-877

We present Spline-based Convolutional Neural Networks (SplineCNNs), a variant of deep neural networks for irregular structured and geometric input, e.g., graphs or meshes. Our main contribution is a novel convolution operator based on B-splines, that makes the computation time independent from the kernel size due to the local support property of the B-spline basis functions. As a result, we obtain a generalization of the traditional CNN convolution operator by using continuous kernel functions parametrized by a fixed number of trainable weights. In contrast to related approaches that filter in the spectral domain, the proposed method aggregates features purely in the spatial domain. In addition, SplineCNN allows entire end-to-end training of deep architectures, using only the geometric structure as input, instead of handcrafted feature descriptors. For validation, we apply our method on tasks from the fields of image graph classification, shape correspondence and graph node classification, and show that it outperforms or pars state-of-the-art approaches while being significantly faster and having favorable properties like domain-independence. Our source code is available on GitHub.

****

GAGAN: Geometry-Aware Generative Adversarial Networks

Jean Kossaifi, Linh Tran, Yannis Panagakis, Maja Pantic; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 878-887

Deep generative models learned through adversarial training have become increasingly popular for their ability to generate naturalistic image textures. However, aside from their texture, the visual appearance of objects is significantly influenced by their shape geometry; information which is not taken into account by existing generative models. This paper introduces the Geometry-Aware Generative Adversarial Networks (GAGAN) for incorporating geometric information into the image generation process. Specifically, in GAGAN the generator samples latent variables from the probability space of a statistical shape model. By mapping the output of the generator to a canonical coordinate frame through a differentiable geometric transformation, we enforce the geometry of the objects and add an implicit connection from the prior to the generated object. Experimental results on face generation indicate that the GAGAN can generate realistic images of faces with arbitrary facial attributes such as facial expression, pose, and morphology, that are of better quality than current GAN-based methods. Our method can be used to augment any existing GAN architecture and improve the quality of the images generated.

****

On the Robustness of Semantic Segmentation Models to Adversarial Attacks

Anurag Arnab, Ondrej Miksik, Philip H.S. Torr; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 888-897

Deep Neural Networks (DNNs) have been demonstrated to perform exceptionally well on most recognition tasks such as image classification and segmentation. However, they have also been shown to be vulnerable to adversarial examples. This phenomenon has recently attracted a lot of attention but it has not been extensively studied on multiple, large-scale datasets and complex tasks such as semantic segmentation which often require more specialised networks with additional components such as CRFs, dilated convolutions, skip-connections and multiscale processing. In this paper, we present what to our knowledge is the first rigorous evaluation of adversarial attacks on modern semantic segmentation models, using two large-scale datasets. We analyse the effect of different network architectures, model capacity and multiscale processing, and show that many observations made on the task of classification do not always transfer to this more complex task. Furthermore, we show how mean-field inference in deep structured models and multiscale processing naturally implement recently proposed adversarial defenses. Our observations will aid future efforts in understanding and defending against adversarial examples. Moreover, in the shorter term, we show which segmentation models should currently be preferred in safety-critical applications due to their inherent robustness.

****

Feedback-Prop: Convolutional Neural Network Inference Under Partial Evidence

Tianlu Wang, Kota Yamaguchi, Vicente Ordonez; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 898-907

We propose an inference procedure for deep convolutional neural networks (CNNs) when partial evidence is available. Our method consists of a general feedback-based propagation approach (feedback-prop) that boosts the prediction accuracy for an arbitrary set of unknown target labels when the values for a non-overlapping arbitrary set of target labels are known. We show that existing models trained in a multi-label or multi-task setting can readily take advantage of feedback-prop without any retraining or fine-tuning. Our feedback-prop inference procedure is general, simple, reliable, and works on different challenging visual recognition tasks. We present two variants of feedback-prop based on layer-wise and residual iterative updates. We experiment using several multi-task models and show that feedback-prop is effective in all of them. Our results unveil a previously unreported but interesting dynamic property of deep CNNs. We also present an associated technical approach that takes advantage of this property for inference under partial evidence in general visual recognition tasks.

****

Super-Resolving Very Low-Resolution Face Images With Supplementary Attributes

Xin Yu, Basura Fernando, Richard Hartley, Fatih Porikli; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 908-917

Given a tiny face image, conventional face hallucination methods aim to super-resolve its high-resolution (HR) counterpart by learning a mapping from an exemplar dataset. Since a low-resolution (LR) input patch may correspond to many HR candidate patches, this ambiguity may lead to erroneous HR facial details and thus distorts final results, such as gender reversal. An LR input contains low-frequency facial components of its HR version while its residual face image defined as the difference between the HR ground-truth and interpolated LR images contains the missing high-frequency facial details. We demonstrate that supplementing residual images or feature maps with facial attribute information can significantly reduce the ambiguity in face super-resolution. To explore this idea, we develop an attribute-embedded upsampling network, which consists of an upsampling network and a discriminative network. The upsampling network is composed of an autoencoder with skip-connections, which incorporates facial attribute vectors into the residual features of LR inputs at the bottleneck of the autoencoder and deconvolutional layers used for upsampling. The discriminative network is designed to examine whether super-resolved faces contain the desired attributes or not and then its loss is used for updating the upsampling network. In this manner, we can super-resolve tiny unaligned (16$■imes$16 pixels) face images with a large upscaling factor of 8$■imes$ while reducing the uncertainty of one-to-many mappings significantly. By conducting extensive evaluations on a large-scale dataset, we demonstrate that our method achieves superior face hallucination results and outperforms the state-of-the-art.

****

Frustum PointNets for 3D Object Detection From RGB-D Data

Charles R. Qi, Wei Liu, Chenxia Wu, Hao Su, Leonidas J. Guibas; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 918-927

In this work, we study 3D object detection from RGB-D data in both indoor and outdoor scenes. While previous methods focus on images or 3D voxels, often obscuring natural 3D patterns and invariances of 3D data, we directly operate on raw point clouds by popping up RGB-D scans. However, a key challenge of this approach is how to efficiently localize objects in point clouds of large-scale scenes (region proposal). Instead of solely relying on 3D proposals, our method leverages both mature 2D object detectors and advanced 3D deep learning for object localization, achieving efficiency as well as high recall for even small objects. Benefited from learning directly in raw point clouds, our method is also able to precisely estimate 3D bounding boxes even under strong occlusion or with very sparse points. Evaluated on KITTI and SUN RGB-D 3D detection benchmarks, our method outperforms the state of the art by remarkable margins while having real-time capability.

****

W2F: A Weakly-Supervised to Fully-Supervised Framework for Object Detection

Yongqiang Zhang, Yancheng Bai, Mingli Ding, Yongqiang Li, Bernard Ghanem; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 928-936

Weakly-supervised object detection has attracted much attention lately, since it does not require bounding box annotations for training. Although significant progress has also been made, there is still a large gap in performance between weakly-supervised and fully-supervised object detection. Recently, some works use pseudo ground-truths which are generated by a weakly-supervised detector to train a supervised detector. Such approaches incline to find the most representative parts of objects, and only seek one ground-truth box per class even though many same-class instances exist. To overcome these issues, we propose a weakly-supervised to fully-supervised framework, where a weakly-supervised detector is implemented using multiple instance learning. Then, we propose a pseudo ground-truth excavation (PGE) algorithm to find the pseudo ground-truth of each instance in the image. Moreover, the pseudo ground-truth adaptation (PGA) algorithm is designe

d to further refine the pseudo ground-truths from PGE. Finally, we use these pseudo ground-truths to train a fully-supervised detector. Extensive experiments on the challenging PASCAL VOC 2007 and 2012 benchmarks strongly demonstrate the effectiveness of our framework. We obtain 52.4% and 47.8% mAP on VOC2007 and VOC2012 respectively, a significant improvement over previous state-of-the-art methods.
****
3D Object Detection With Latent Support Surfaces
Zhile Ren, Erik B. Sudderth; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 937-946
We develop a 3D object detection algorithm that uses latent support surfaces to capture contextual relationships in indoor scenes. Existing 3D representations for RGB-D images capture the local shape and appearance of object categories, but have limited power to represent objects with different visual styles. The detection of small objects is also challenging because the search space is very large in 3D scenes. However, we observe that much of the shape variation within 3D object categories can be explained by the location of a latent support surface, and smaller objects are often supported by larger objects. Therefore, we explicitly use latent support surfaces to better represent the 3D appearance of large objects, and provide contextual cues to improve the detection of small objects. We evaluate our model with 19 object categories from the SUN RGB-D database, and demonstrate state-of-the-art performance.
****
Towards Faster Training of Global Covariance Pooling Networks by Iterative Matrix Square Root Normalization
Peihua Li, Jiangtao Xie, Qilong Wang, Zilin Gao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 947-955
Global covariance pooling in convolutional neural networks has achieved impressive improvement over the classical first-order pooling. Recent works have shown matrix square root normalization plays a central role in achieving state-of-the-art performance. However, existing methods depend heavily on eigendecomposition (EIG) or singular value decomposition (SVD), suffering from inefficient training due to limited support of EIG and SVD on GPU. Towards addressing this problem, we propose an iterative matrix square root normalization method for fast end-to-end training of global covariance pooling networks. At the core of our method is a meta-layer designed with loop-embedded directed graph structure. The meta-layer consists of three consecutive nonlinear structured layers, which perform pre-normalization, coupled matrix iteration and post-compensation, respectively. Our method is much faster than EIG or SVD based ones, since it involves only matrix multiplications, suitable for parallel implementation on GPU. Moreover, the proposed network with ResNet architecture can converge in much less epochs, further accelerating network training. On large-scale ImageNet, we achieve competitive performance superior to existing counterparts. By finetuning our models pre-trained on ImageNet, we establish state-of-the-art results on three challenging fine-grained benchmarks. The source code and network models will be available at http://www.peihuali.org/iSQRT-COV.
****
Recurrent Scene Parsing With Perspective Understanding in the Loop
Shu Kong, Charless C. Fowlkes; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 956-965
Objects may appear at arbitrary scales in perspective images of a scene, posing a challenge for recognition systems that process images at a fixed resolution. We propose a depth-aware gating module that adaptively selects the pooling field size in a convolutional network architecture according to the object scale (inversely proportional to the depth) so that small details are preserved for distant objects while larger receptive fields are used for those nearby. The depth gating signal is provided by stereo disparity or estimated directly from monocular input. We integrate this depth-aware gating into a recurrent convolutional neural network to perform semantic segmentation. Our recurrent module iteratively efines the segmentation results, leveraging the depth and semantic predictions from

the previous iterations. Through extensive experiments on four popular large-scale datasets, we demonstrate this approach achieves competitive semantic segmentation performance with a model which is substantially more compact. We carry out extensive analysis of this architecture including variants that operate on monocular RGB but use depth as side-information during training, unsupervised gating as a generic attentional mechanism, and multi-resolution gating. We find that gated pooling for joint semantic segmentation and depth yields state-of-the-art results for quantitative monocular depth estimation.

****

Improving Occlusion and Hard Negative Handling for Single-Stage Pedestrian Detectors

Junhyug Noh, Soochan Lee, Beomsu Kim, Gunhee Kim; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 966-974

We propose methods of addressing two critical issues of pedestrian detection: (i) occlusion of target objects as false negative failure, and (ii) confusion with hard negative examples like vertical structures as false positive failure. Our solutions to these two problems are general and flexible enough to be applicable to any single-stage detection models. We implement our methods into four state-of-the-art single-stage models, including SqueezeDet+, YOLOv2, SSD, and DSSD. We empirically validate that our approach indeed improves the performance of those four models on Caltech pedestrian and CityPersons dataset. Moreover, in some heavy occlusion settings, our approach achieves the best reported performance. Specifically, our two solutions are as follows. For better occlusion handling, we update the output tensors of single-stage models so that they include the prediction of part confidence scores, from which we compute a final occlusion-aware detection score. For reducing confusion with hard negative examples, we introduce average grid classifiers as post-refinement classifiers, trainable in an end-to-end fashion with little memory and time overhead (e.g. increase of 1--5 MB in memory and 1--2 ms in inference time).

****

Learning to Act Properly: Predicting and Explaining Affordances From Images

Ching-Yao Chuang, Jiaman Li, Antonio Torralba, Sanja Fidler; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 975-983

We address the problem of affordance reasoning in diverse scenes that appear in the real world. Affordances relate the agent's actions to their effects when taken on the surrounding objects. In our work, we take the egocentric view of the scene, and aim to reason about action-object affordances that respect both the physical world as well as the social norms imposed by the society. We also aim to teach artificial agents why some actions should not be taken in certain situations, and what would likely happen if these actions would be taken. We collect a new dataset that builds upon ADE20k, referred to as ADE-Affordance, which containing annotations enabling such rich visual reasoning. We propose a model that exploits Graph Neural Networks to propagate contextual information from the scene in order to perform detailed affordance reasoning about each object. Our model is showcased through various ablation studies, pointing to successes and challenges in this complex task.

****

Pointwise Convolutional Neural Networks

Binh-Son Hua, Minh-Khoi Tran, Sai-Kit Yeung; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 984-993

Deep learning with 3D data such as reconstructed point clouds and CAD models has received great research interests recently. However, the capability of using point clouds with convolutional neural network has been so far not fully explored. In this paper, we present a convolutional neural network for semantic segmentation and object recognition with 3D point clouds. At the core of our network is pointwise convolution, a new convolution operator that can be applied at each point of a point cloud. Our fully convolutional network design, while being surprisingly simple to implement, can yield competitive accuracy in both semantic segmentation and object recognition task.

****

Image-Image Domain Adaptation With Preserved Self-Similarity and Domain-Dissimilarity for Person Re-Identification

Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, Jianbin Jiao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 994-1003

Person re-identification (re-ID) models trained on one domain often fail to generalize well to another. In our attempt, we present a ``learning via translation'' framework. In the baseline, we translate the labeled images from source to target domain in an unsupervised manner. We then train re-ID models with the translated images by supervised methods. Yet, being an essential part of this framework, unsupervised image-image translation suffers from the information loss of source-domain labels during translation. Our motivation is two-fold. First, for each image, the discriminative cues contained in its ID label should be maintained after translation. Second, given the fact that two domains have entirely different persons, a translated image should be dissimilar to any of the target IDs. To this end, we propose to preserve two types of unsupervised similarities, 1) self-similarity of an image before and after translation, and 2) domain-dissimilarity of a translated source image and a target image. Both constraints are implemented in the similarity preserving generative adversarial network (SPGAN) which consists of an Siamese network and a CycleGAN. Through domain adaptation experiment, we show that images generated by SPGAN are more suitable for domain adaptation and yield consistent and competitive re-ID accuracy on two large-scale datasets.

****

A Generative Adversarial Approach for Zero-Shot Learning From Noisy Texts

Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Xi Peng, Ahmed Elgammal; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1004-1013

Most existing zero-shot learning methods consider the problem as a visual semantic embedding one. Given the demonstrated capability of Generative Adversarial Networks(GANs) to generate images, we instead leverage GANs to imagine unseen categories from text descriptions and hence recognize novel classes with no examples being seen. Specifically, we propose a simple yet effective generative model that takes as input noisy text descriptions about an unseen class (e.g.Wikipedia articles) and generates synthesized visual features for this class. With added pseudo data, zero-shot learning is naturally converted to a traditional classification problem. Additionally, to preserve the inter-class discrimination of the generated features, a visual pivot regularization is proposed as an explicit supervision. Unlike previous methods using complex engineered regularizers, our approach can suppress the noise well without additional regularization. Empirically, we show that our method consistently outperforms the state of the art on the largest available benchmarks on Text-based Zero-shot Learning.

****

Tensorize, Factorize and Regularize: Robust Visual Relationship Learning

Seong Jae Hwang, Sathya N. Ravi, Zirui Tao, Hyunwoo J. Kim, Maxwell D. Collins, Vikas Singh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1014-1023

Visual relationships provide higher-level information of objects and their relations in an image – this enables a semantic understanding of the scene and helps downstream applications. Given a set of localized objects in some training data, visual relationship detection seeks to detect the most likely “relationship” between objects in a given image. While the specific objects may be well represented in training data, their relationships may still be infrequent. The empirical distribution obtained from seeing these relationships in a dataset does not model the underlying distribution well — a serious issue for most learning methods. In this work, we start from a simple multi-relational learning model, which in principle, offers a rich formalization for deriving a strong prior for learning visual relationships. While the inference problem for deriving the regularizer is challenging, our main technical contribution is to show how adapting recent res ults in numerical linear algebra lead to efficient algorithms for a factorization scheme that yields highly informative priors. The factorization provides sample size bounds for inference (under mild conditions) for the underlying [[object, predicate, object]] relationship learning task on its own and surprisingly outperforms (in some cases) existing methods even without utilizing visual features. Then, when integrated with an end to-end architecture for visual relationship detection leveraging image data, we substantially improve the state-of-the-art.

****

Transductive Unbiased Embedding for Zero-Shot Learning

Jie Song, Chengchao Shen, Yezhou Yang, Yang Liu, Mingli Song; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1024-1033

Most existing Zero-Shot Learning (ZSL) methods have the strong bias problem, in which instances of unseen (target) classes tend to be categorized as one of the seen (source) classes. So they yield poor performance after being deployed in the generalized ZSL settings. In this paper, we propose a straightforward yet effective method named Quasi-Fully Supervised Learning (QFSL) to alleviate the bias problem. Our method follows the way of transductive learning, which assumes that both the labeled source images and unlabeled target images are available for training. In the semantic embedding space, the labeled source images are mapped to several fixed points specified by the source categories, and the unlabeled target images are forced to be mapped to other points specified by the target categories. Experiments conducted on AwA2, CUB and SUN datasets demonstrate that our method outperforms existing state-of-the-art approaches by a huge margin of 9.3~24.5% following generalized ZSL settings, and by a large margin of 0.2~16.2% following conventional ZSL settings.

****

Hierarchical Novelty Detection for Visual Object Recognition

Kibok Lee, Kimin Lee, Kyle Min, Yuting Zhang, Jinwoo Shin, Honglak Lee; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1034-1042

Deep neural networks have achieved impressive success in large-scale visual object recognition tasks with a predefined set of classes. However, recognizing objects of novel classes unseen during training still remains challenging. The problem of detecting such novel classes has been addressed in the literature, but most prior works have focused on providing simple binary or regressive decisions, e.g., the output would be "known," "novel," or corresponding confidence intervals. In this paper, we study more informative novelty detection schemes based on a hierarchical classification framework. For an object of a novel class, we aim for finding its closest super class in the hierarchical taxonomy of known classes. To this end, we propose two different approaches termed top-down and flatten methods, and their combination as well. The essential ingredients of our methods are confidence-calibrated classifiers, data relabeling, and the leave-one-out strategy for modeling novel classes under the hierarchical taxonomy. Furthermore, our method can generate a hierarchical embedding that leads to improved generalized zero-shot learning performance in combination with other commonly-used semantic embeddings.

****

Zero-Shot Visual Recognition Using Semantics-Preserving Adversarial Embedding Networks

Long Chen, Hanwang Zhang, Jun Xiao, Wei Liu, Shih-Fu Chang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1043-1052

We propose a novel framework called Semantics-Preserving Adversarial Embedding Network (SP-AEN) for zero-shot visual recognition (ZSL), where test images and their classes are both unseen during training. SP-AEN aims to tackle the inherent problem — semantic loss — in the prevailing family of embedding-based ZSL, where some semantics would be discarded during training if they are non-discriminative for training classes, but could become critical for recognizing test classes. Specifically, SP-AEN prevents the semantic loss by introducing an independent vi sual-to-semantic space embedder which disentangles the semantic space into two subspaces for the two arguably conflicting objectives: classification and reconstruction. Through adversarial learning of the two subspaces, SP-AEN can transfer the semantics from the reconstructive subspace to the discriminative one, accomplishing the improved zero-shot recognition of unseen classes. Comparing with prior works, SP-AEN can not only improve classification but also generate photo-realistic images, demonstrating the effectiveness of semantic preservation. On four popular benchmarks: CUB, AWA, SUN and aPY, SP-AEN considerably outperforms other state-of-the-art methods by an absolute performance difference of 12.2%, 9.3%, 4.0%, and 3.6% in terms of harmonic mean values.

****

Learning Rich Features for Image Manipulation Detection

Peng Zhou, Xintong Han, Vlad I. Morariu, Larry S. Davis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1053-1061

Image manipulation detection is different from traditional semantic object detection because it pays more attention to tampering artifacts than to image content, which suggests that richer features need to be learned. We propose a two-stream Faster R-CNN network and train it end-to- end to detect the tampered regions given a manipulated image. One of the two streams is an RGB stream whose purpose is to extract features from the RGB image input to find tampering artifacts like strong contrast difference, unnatural tampered boundaries, and so on. The other is a noise stream that leverages the noise features extracted from a steganalysis rich model filter layer to discover the noise inconsistency between authentic and tampered regions. We then fuse features from the two streams through a bilinear pooling layer to further incorporate spatial co-occurrence of these two modalities. Experiments on four standard image manipulation datasets demonstrate that our two-stream framework outperforms each individual stream, and also achieves state-of-the-art performance compared to alternative methods with robustness to resizing and compression.

****

Human Semantic Parsing for Person Re-Identification

Mahdi M. Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E. Kamasak, Mubarak Shah; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1062-1071

Person re-identification is a challenging task mainly due to factors such as background clutter, pose, illumination and camera point of view variations. These elements hinder the process of extracting robust and discriminative representations, hence preventing different identities from being successfully distinguished. To improve the representation learning, usually local features from human body parts are extracted. However, the common practice for such a process has been based on bounding box part detection. In this paper, we propose to adopt human semantic parsing which, due to its pixel-level accuracy and capability of modeling arbitrary contours, is naturally a better alternative. Our proposed SPReID integrates human semantic parsing in person re-identification and not only considerably outperforms its counter baseline, but achieves state-of-the-art performance. We also show that, by employing a simple yet effective training strategy, standard popular deep convolutional architectures such as Inception-V3 and ResNet-152, with no modification, while operating solely on full image, can dramatically outperform current state-of-the-art. Our proposed methods improve state-of-the-art person re-identification on: Market-1501 by ~17% in mAP and ~6% in rank-1, CUHK03 by ~4% in rank-1 and DukeMTMC-reID by ~24% in mAP and ~10% in rank-1.

****

Stacked Latent Attention for Multimodal Reasoning

Haoqi Fan, Jiatong Zhou; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1072-1080

Attention has shown to be a pivotal development in deep learning and has been used for a multitude of multimodal learning tasks such as visual question answering and image captioning. In this work, we pinpoint the potential limitations to the design of a traditional attention model. We identify that 1) current attentio n mechanisms discard the latent information from intermediate reasoning, losing the positional information already captured by the attention heatmaps and 2) stacked attention, a common way to improve spatial reasoning, may have suboptimal performance because of the vanishing gradient problem. We introduce a novel attention architecture to address these problems, in which all spatial configuration information contained in the intermediate reasoning process is retained in a pathway of convolutional layers. We show that this new attention leads to substantial improvements in multiple multimodal reasoning tasks, including achieving single model performance without using external knowledge comparable to the state-of-the-art on the VQA dataset, as well as clear gains for the image captioning task.

****

R-FCN-3000 at 30fps: Decoupling Detection and Classification

Bharat Singh, Hengduo Li, Abhishek Sharma, Larry S. Davis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1081-1090

We propose a modular approach towards large-scale real-time object detection by decoupling objectness detection and classification. We exploit the fact that many object classes are visually similar and share parts. Thus, a universal objectness detector can be learned for class-agnostic object detection followed by fine-grained classification using a (non)linear classifier. Our approach is a modification of the R-FCN architecture to learn shared filters for performing localization across different object classes. We trained a detector for 3000 object classes, called R-FCN-3000, that obtains an mAP of 34.9% on the ImageNet detection dataset. It outperforms YOLO-9000 by 18% while processing 30 images per second. We also show that the objectness learned by R-FCN-3000 generalizes to novel classes and the performance increases with the number of training object classes - supporting the hypothesis that it is possible to learn a universal objectness detector.

****

CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes

Yuhong Li, Xiaofan Zhang, Deming Chen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1091-1100

We propose a network for Congested Scene Recognition called CSRNet to provide a data-driven and deep learning method that can understand highly congested scenes and perform accurate count estimation as well as present high-quality density maps. The proposed CSRNet is composed of two major components: a convolutional neural network (CNN) as the front-end for 2D feature extraction and a dilated CNN for the back-end, which uses dilated kernels to deliver larger reception fields and to replace pooling operations. CSRNet is an easy-trained model because of its pure convolutional structure. We demonstrate CSRNet on four datasets (ShanghaiTech dataset, the UCF_CC_50 dataset, the WorldEXPO'10 dataset, and the UCSD dataset) and we deliver the state-of-the-art performance. In the ShanghaiTech Part_B dataset, CSRNet achieves 47.3% lower Mean Absolute Error (MAE) than the previous state-of-the-art method. We extend the targeted applications for counting other objects, such as the vehicle in TRANCOS dataset. Results show that CSRNet significantly improves the output quality with 15.4% lower MAE than the previous state-of-the-art approach.

****

Revisiting Knowledge Transfer for Training Object Class Detectors

Jasper Uijlings, Stefan Popov, Vittorio Ferrari; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1101-1110

We propose to revisit knowledge transfer for training object detectors on target classes from weakly supervised training images, helped by a set of source classes with bounding-box annotations. We present a unified knowledge transfer framework based on training a single neural network multi-class object detector over all source classes, organized in a semantic hierarchy. This generates proposals with scores at multiple levels in the hierarchy, which we use to explore knowledge transfer over a broad range of generality, ranging from class-specific (bycicl e to motorbike) to class-generic (objectness to any class). Experiments on the 200 object classes in the ILSVRC 2013 detection dataset show that our technique (1) leads to much better performance on the target classes (70.3% CorLoc, 36.9% mAP) than a weakly supervised baseline which uses manually engineered objectness [11] (50.5% CorLoc, 25.4% mAP). (2) delivers target object detectors reaching 80% of the mAP of their fully supervised counterparts. (3) outperforms the best reported transfer learning results on this dataset (+41% CorLoc and +3% mAP over [18, 46], +16.2% mAP over [32]). Moreover, we also carry out several across-dataset knowledge transfer experiments [27, 24, 35] and find that (4) our technique outperforms the weakly supervised baseline in all dataset pairs by 1.5 × −1.9×, establishing its general applicability.

****

Deep Sparse Coding for Invariant Multimodal Halle Berry Neurons

Edward Kim, Darryl Hannan, Garrett Kenyon; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1111-1120

Deep feed-forward convolutional neural networks (CNNs) have become ubiquitous in virtually all machine learning and computer vision challenges; however, advancements in CNNs have arguably reached an engineering saturation point where incremental novelty results in minor performance gains. Although there is evidence that object classification has reached human levels on narrowly defined tasks, for general applications, the biological visual system is far superior to that of any computer. Research reveals there are numerous missing components in feed-forward deep neural networks that are critical in mammalian vision. The brain does not work solely in a feed-forward fashion, but rather all of the neurons are in competition with each other; neurons are integrating information in a bottom up and top down fashion and incorporating expectation and feedback in the modeling process. Furthermore, our visual cortex is working in tandem with our parietal lobe, integrating sensory information from various modalities. In our work, we sought to improve upon the standard feed-forward deep learning model by augmenting them with biologically inspired concepts of sparsity, top down feedback, and lateral inhibition. We define our model as a sparse coding problem using hierarchical layers. We solve the sparse coding problem with an additional top down feedback error driving the dynamics of the neural network. While building and observing the behavior of our model, we were fascinated that multimodal, invariant neurons naturally emerged that mimicked, "Halle Berry neurons" found in the human brain. These neurons trained in our sparse model learned to respond to high level concepts from multiple modalities, which is not the case with a standard feed-forward autoencoder. Furthermore, our sparse representation of multimodal signals demonstrates qualitative and quantitative superiority to the standard feed-forward joint embedding in common vision and machine learning tasks.

****

On the Convergence of PatchMatch and Its Variants

Thibaud Ehret, Pablo Arias; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1121-1129

Many problems in image/video processing and computer vision require the computation of a dense k-nearest neighbor field (k-NNF) between two images. For each patch in a query image, the k-NNF determines the positions of the k most similar patches in a database image. With the introduction of the PatchMatch algorithm, Barnes et al. demonstrated that this large search problem can be approximated efficiently by collaborative search methods that exploit the local coherency of image patches. After its introduction, several variants of the original PatchMatch algorithm have been proposed, some of them reducing the computational time by two orders of magnitude. In this work we propose a theoretical framework for the analysis of PatchMatch and its variants, and apply it to derive bounds on their covergence rate. We consider a generic PatchMatch algorithm from which most specific instances found in the literature can be derived as particular cases. We also derive more specific bounds for two of these particular cases: the original PatchMatch and Coherency Sensitive Hashing. The proposed bounds are validated by contrasting them to the convergence observed in practice.

****

Rethinking the Faster R-CNN Architecture for Temporal Action Localization

Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A. Ross, Jia Deng, Rahul Sukthankar; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1130-1139

We propose TAL-Net, an improved approach to temporal action localization in video that is inspired by the Faster R-CNN object detection framework. TAL-Net addresses three key shortcomings of existing approaches: (1) we improve receptive field alignment using a multi-scale architecture that can accommodate extreme variation in action durations; (2) we better exploit the temporal context of actions for both proposal generation and action classification by appropriately extending receptive fields; and (3) we explicitly consider multi-stream feature fusion and demonstrate that fusing motion late is important. We achieve state-of-the-art performance for both action proposal and localization on THUMOS'14 detection benchmark and competitive performance on ActivityNet challenge.

****

MoNet: Deep Motion Exploitation for Video Object Segmentation

Huaxin Xiao, Jiashi Feng, Guosheng Lin, Yu Liu, Maojun Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1140-1148

In this paper, we propose a novel MoNet model to deeply exploit motion cues for boosting video object segmentation performance from two aspects, i.e., frame representation learning and segmentation refinement. Concretely, MoNet exploits computed motion cue (i.e., optical flow) to reinforce the representation of the target frame by aligning and integrating representations from its neighbors. The new representation provides valuable temporal contexts for segmentation and improves robustness to various common contaminating factors, e.g., motion blur, appearance variation and deformation of video objects. Moreover, MoNet exploits motion inconsistency and transforms such motion cue into foreground/background prior to eliminate distraction from confusing instances and noisy regions. By introducing a distance transform layer, MoNet can effectively separate motion-inconstant instances/regions and thoroughly refine segmentation results. Integrating the proposed two motion exploitation components with a standard segmentation network, MoNet provides new state-of-the-art performance on three competitive benchmark datasets.

****

Video Representation Learning Using Discriminative Pooling

Jue Wang, Anoop Cherian, Fatih Porikli, Stephen Gould; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1149-1158

Popular deep models for action recognition in videos generate independent predictions for short clips, which are then pooled heuristically to assign an action label to the full video segment. As not all frames may characterize the underlying action---indeed, many are common across multiple actions---pooling schemes that impose equal importance on all frames might be unfavorable. In an attempt to tackle this problem, we propose discriminative pooling, based on the notion that among the deep features generated on all short clips, there is at least one that characterizes the action. To this end, we learn a (nonlinear) hyperplane that separates this unknown, yet discriminative, feature from the rest. Applying multiple instance learning in a large-margin setup, we use the parameters of this separating hyperplane as a descriptor for the full video segment. Since these parameters are directly related to the support vectors in a max-margin framework, they serve as robust representations for pooling of the features. We formulate a joint objective and an efficient solver that learns these hyperplanes per video and the corresponding action classifiers over the hyperplanes. Our pooling scheme is end-to-end trainable within a deep framework. We report results from experiments on three benchmark datasets spanning a variety of challenges and demonstrate state-of-the-art performance across these tasks.

****

Recognizing Human Actions as the Evolution of Pose Estimation Maps

Mengyuan Liu, Junsong Yuan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1159-1168

Most video-based action recognition approaches choose to extract features from the whole video to recognize actions. The cluttered background and non-action motions limit the performances of these methods, since they lack the explicit modeling of human body movements. With recent advances of human pose estimation, this work presents a novel method to recognize human action as the evolution of pose estimation maps. Instead of relying on the inaccurate human poses estimated from videos, we observe that pose estimation maps, the byproduct of pose estimation, preserve richer cues of human body to benefit action recognition. Specifically, the evolution of pose estimation maps can be decomposed as an evolution of heatmaps, e.g., probabilistic maps, and an evolution of estimated 2D human poses, which denote the changes of body shape and body pose, respectively. Considering the sparse property of heatmap, we develop spatial rank pooling to aggregate the evolution of heatmaps as a body shape evolution image. As body shape evolution image does not differentiate body parts, we design body guided sampling to aggregate the evolution of poses as a body pose evolution image. The complementary properties between both types of images are explored by deep convolutional neural networks to predict action label. Experiments on NTU RGB+D, UTD-MHAD and PennAction datasets verify the effectiveness of our method, which outperforms most state-of-the-art methods.

****

Video Person Re-Identification With Competitive Snippet-Similarity Aggregation and Co-Attentive Snippet Embedding

Dapeng Chen, Hongsheng Li, Tong Xiao, Shuai Yi, Xiaogang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1169-1178

In this paper, we address video-based person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. Our approach divides long person sequences into multiple short video snippets and aggregates the top-ranked snippet similarities for sequence-similarity estimation. With this strategy, the intra-person visual variation of each sample could be minimized for similarity estimation, while the diverse appearance and temporal information are maintained. The snippet similarities are estimated by a deep neural network with a novel temporal co-attention for snippet embedding. The attention weights are obtained based on a query feature, which is learned from the whole probe snippet by an LSTM network, making the resulting embeddings less affected by noisy frames. The gallery snippet shares the same query feature with the probe snippet. Thus the embedding of gallery snippet can present more relevant features to compare with the probe snippet, yielding more accurate snippet similarity. Extensive ablation studies verify the effectiveness of competitive snippet-similarity aggregation as well as the temporal co-attentive embedding. Our method significantly outperforms the current state-of-the-art approaches on multiple datasets.

****

Mask-Guided Contrastive Attention Model for Person Re-Identification

Chunfeng Song, Yan Huang, Wanli Ouyang, Liang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1179-1188

Person Re-identification (ReID) is an important yet challenging task in computer vision. Due to the diverse background clutters, variations on viewpoints and body poses, it is far from solved. How to extract discriminative and robust features invariant to background clutters is the core problem. In this paper, we first introduce the binary segmentation masks to construct synthetic RGB-Mask pairs as inputs, then we design a mask-guided contrastive attention model (MGCAM) to learn features separately from the body and background regions. Moreover, we propose a novel region-level triplet loss to restrain the features learnt from different regions, i.e., pulling the features from the full image and body region close, whereas pushing the features from backgrounds away. We may be the first one to successfully introduce the binary mask into person ReID task and the first one to propose region-level contrastive learning. We evaluate the proposed method on three public datasets, including MARS, Market-1501 and CUHK03. Extensive experimental results show that the proposed method is effective and achieves the stat e-of-the-art results. Mask and code will be released upon request.

****

Blazingly Fast Video Object Segmentation With Pixel-Wise Metric Learning

Yuhua Chen, Jordi Pont-Tuset, Alberto Montes, Luc Van Gool; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1189-1198

This paper tackles the problem of video object segmentation, given some user annotation which indicates the object of interest. The problem is formulated as pixel-wise retrieval in a learned embedding space: we embed pixels of the same object instance into the vicinity of each other, using a fully convolutional network trained by a modified triplet loss as the embedding model. Then the annotated pixels are set as reference and the rest of the pixels are classified using a nearest-neighbor approach. The proposed method supports different kinds of user input such as segmentation mask in the first frame (semi-supervised scenario), or a sparse set of clicked points (interactive scenario). In the semi-supervised scenario, we achieve results competitive with the state of the art but at a fraction of computation cost (275 milliseconds per frame). In the interactive scenario where the user is able to refine their input iteratively, the proposed method provides instant response to each input, and reaches comparable quality to competing methods with much less interaction.

****

Learning to Compare: Relation Network for Few-Shot Learning

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H.S. Torr, Timothy M. Hospedales; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1199-1208

We present a conceptually simple, flexible, and general framework for few-shot learning, where a classifier must learn to recognise new classes given only few examples from each. Our method, called the Relation Network (RN), is trained end-to-end from scratch. During meta-learning, it learns to learn a deep distance metric to compare a small number of images within episodes, each of which is designed to simulate the few-shot setting. Once trained, a RN is able to classify images of new classes by computing relation scores between query images and the few examples of each new class without further updating the network. Besides providing improved performance on few-shot learning, our framework is easily extended to zero-shot learning. Extensive experiments on five benchmarks demonstrate that our simple approach provides a unified and effective approach for both of these two tasks.

****

COCO-Stuff: Thing and Stuff Classes in Context

Holger Caesar, Jasper Uijlings, Vittorio Ferrari; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1209-1218

Semantic classes can be either things (objects with a well-defined shape, e.g. car, person) or stuff (amorphous background regions, e.g. grass, sky). While lots of classification and detection works focus on thing classes, less attention has been given to stuff classes. Nonetheless, stuff classes are important as they allow to explain important aspects of an image, including (1) scene type; (2) which thing classes are likely to be present and their location (through contextual reasoning); (3) physical attributes, material types and geometric properties of the scene. To understand stuff and things in context we introduce COCO-Stuff, which augments all 164K images of the COCO 2017 dataset with pixel-wise annotations for 91 stuff classes. We introduce an efficient stuff annotation protocol based on superpixels, which leverages the original thing annotations. We quantify the speed versus quality trade-off of our protocol and explore the relation between annotation time and boundary complexity. Furthermore, we use COCO-Stuff to analyze: (a) the importance of stuff and thing classes in terms of their surface cover and how frequently they are mentioned in image captions; (b) the spatial relations between stuff and things, highlighting the rich contextual relations that make our dataset unique; (c) the performance of a modern semantic segmentation method on stuff and thing classes, and whether stuff is easier to segment than things.

****

Image Generation From Scene Graphs

Justin Johnson, Agrim Gupta, Li Fei-Fei; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1219-1228

To truly understand the visual world our models should be able not only to recognize images but also generate them. To this end, there has been exciting recent progress on gen- erating images from natural language descriptions. These methods give stunning results on limited domains such as descriptions of birds or flowers, but struggle to faithfully reproduce complex sentences with many objects and rela- tionships. To overcome this limitation we propose a method for generating images from scene graphs, enabling explic- itly reasoning about objects and their relationships. Our model uses graph convolution to process input graphs, com- putes a scene layout by predicting bounding boxes and seg- mentation masks for objects, and converts the layout to an image with a cascaded refinement network. The network is trained adversarially against a pair of discriminators to en- sure realistic outputs. We validate our approach on Visual Genome and COCO-Stuff, where qualitative results, abla- tions, and user studies demonstrate our method’s ability to generate complex images with multiple objects.

****

Deep Cauchy Hashing for Hamming Space Retrieval

Yue Cao, Mingsheng Long, Bin Liu, Jianmin Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1229-1237

Due to its computation efficiency and retrieval quality, hashing has been widely applied to approximate nearest neighbor search for large-scale image retrieval, while deep hashing further improves the retrieval quality by end-to-end representation learning and hash coding. With compact hash codes, Hamming space retrieval enables the most efficient constant-time search that returns data points within a given Hamming radius to each query, by hash table lookups instead of linear scan. However, subject to the weak capability of concentrating relevant images to be within a small Hamming ball due to mis-specified loss functions, existing deep hashing methods may underperform for Hamming space retrieval. This work presents Deep Cauchy Hashing (DCH), a novel deep hashing model that generates compact and concentrated binary hash codes to enable efficient and effective Hamming space retrieval. The main idea is to design a pairwise cross-entropy loss based on Cauchy distribution, which penalizes significantly on similar image pairs with Hamming distance larger than the given Hamming radius threshold. Comprehensive experiments demonstrate that DCH can generate highly concentrated hash codes and yield state-of-the-art Hamming space retrieval performance on three datasets, NUS-WIDE, CIFAR-10, and MS-COCO.

****

Learning to Look Around: Intelligently Exploring Unseen Environments for Unknown Tasks

Dinesh Jayaraman, Kristen Grauman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1238-1247

It is common to implicitly assume access to intelligently captured inputs (e.g., photos from a human photographer), yet autonomously capturing good observations is itself a major challenge. We address the problem of learning to look around: if an agent has the ability to voluntarily acquire new views to observe its environment, how can it learn efficient exploratory behaviors to acquire informative visual observations? We propose a reinforcement learning solution, where the agent is rewarded for actions that reduce its uncertainty about the unobserved portions of its environment. Based on this principle, we develop a recurrent neural network-based approach to perform active completion of panoramic natural scenes and 3D object shapes. Crucially, the learned policies are not tied to any recognition task nor to the particular semantic content seen during training. As a result, 1) the learned "look around" behavior is relevant even for new tasks in unseen environments, and 2) training data acquisition involves no manual labeling. Through tests in diverse settings, we demonstrate that our approach learns useful generic policies that transfer to new unseen tasks and environments.

****

Multi-Scale Location-Aware Kernel Representation for Object Detection

Hao Wang, Qilong Wang, Mingqi Gao, Peihua Li, Wangmeng Zuo; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1248-1257

Although Faster R-CNN and its variants have shown promising performance in object detection, they only exploit simple first order representation of object proposals for final classification and regression. Recent classification methods demonstrate that the integration of high order statistics into deep convolutional neural networks can achieve impressive improvement, but their goal is to model whole images by discarding location information so that they cannot be directly adopted to object detection. In this paper, we make an attempt to exploit high-order statistics in object detection, aiming at generating more discriminative representations for proposals to enhance the performance of detectors. To this end, we propose a novel Multi-scale Location-aware Kernel Representation (MLKP) to capture high-order statistics of deep features in proposals. Our MLKP can be efficiently computed on a modified multi-scale feature map using a low-dimensional polynomial kernel approximation. Moreover, different from existing orderless global representations based on high-order statistics, our proposed MLKP is location retentive and sensitive so that it can be flexibly adopted to object detection. Through integrating into Faster R-CNN schema, the proposed MLKP achieves very competitive performance with state-of-the-art methods, and improves Faster R-CNN by 4.9% (mAP), 4.7% (mAP) and 5.0 (AP at IOU=[0.5:0.05:0.95]) on PASCAL VOC 2007, VOC 2012 and MS COCO benchmarks, respectively. Code is available at: https://github.com/Hwang64/MLKP.

****

Clinical Skin Lesion Diagnosis Using Representations Inspired by Dermatologist Criteria

Jufeng Yang, Xiaoxiao Sun, Jie Liang, Paul L. Rosin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1258-1266

The skin is the largest organ in human body. Around 30%-70% of individuals worldwide have skin related health problems, for whom effective and efficient diagnosis is necessary. Recently, computer aided diagnosis (CAD) systems have been successfully applied to the recognition of skin cancers in dermatoscopic images. However, little work has concentrated on the commonly encountered skin diseases in clinical images captured by easily-accessed cameras or mobile phones. Meanwhile, for a CAD system, the representations of skin lesions are required to be understandable for dermatologists so that the predictions are convincing. To address this problem, we present effective representations inspired by the accepted dermatological criteria for diagnosing clinical skin lesions. We demonstrate that the dermatological criteria are highly correlated with measurable visual components. Accordingly, we design six medical representations considering different criteria for the recognition of skin lesions, and construct a diagnosis system for clinical skin disease images. Experimental results show that the proposed medical representations can not only capture the manifestations of skin lesions effectively, and consistently with the dermatological criteria, but also improve the prediction performance with respect to the state-of-the-art methods based on uninterpretable features.

****

Compare and Contrast: Learning Prominent Visual Differences

Steven Chen, Kristen Grauman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1267-1276

Relative attribute models can compare images in terms of all detected properties or attributes, exhaustively predicting which image is fancier, more natural, and so on without any regard to ordering. However, when humans compare images, certain differences will naturally stick out and come to mind first. These most noticeable differences, or prominent differences, are likely to be described first. In addition, many differences, although present, may not be mentioned at all. In this work, we introduce and model prominent differences, a rich new functionality for comparing images. We collect instance-level annotations of most noticeable differences, and build a model trained on relative attribute features that pr edicts prominent differences for unseen pairs. We test our model on the challenging UT-Zap50K shoes and LFW-10 faces datasets, and outperform an array of baseline methods. We then demonstrate how our prominence model improves two vision tasks, image search and description generation, enabling more natural communication between people and vision systems.

****

Multi-Evidence Filtering and Fusion for Multi-Label Classification, Object Detection and Semantic Segmentation Based on Weakly Supervised Learning

Weifeng Ge, Sibei Yang, Yizhou Yu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1277-1286

Supervised object detection and semantic segmentation require object or even pixel level annotations. When there exist image level labels only, it is challenging for weakly supervised algorithms to achieve accurate predictions. The accuracy achieved by top weakly supervised algorithms is still significantly lower than their fully supervised counterparts. In this paper, we propose a novel weakly supervised curriculum learning pipeline for multi-label object recognition, detection and semantic segmentation. In this pipeline, we first obtain intermediate object localization and pixel labeling results for the training images, and then use such results to train task-specific deep networks in a fully supervised manner. The entire process consists of four stages, including object localization in the training images, filtering and fusing object instances, pixel labeling for the training images, and task-specific network training. To obtain clean object instances in the training images, we propose a novel algorithm for filtering, fusing and classifying object instances collected from multiple solution mechanisms. In this algorithm, we incorporate both metric learning and density-based clustering to filter detected object instances. Experiments show that our weakly supervised pipeline achieves state-of-the-art results in multi-label image classification as well as weakly supervised object detection and very competitive results in weakly supervised semantic segmentation on MS-COCO, PASCAL VOC 2007 and PASCAL VOC 2012.

****

HashGAN: Deep Learning to Hash With Pair Conditional Wasserstein GAN

Yue Cao, Bin Liu, Mingsheng Long, Jianmin Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1287-1296

Deep learning to hash improves image retrieval performance by end-to-end representation learning and hash coding from training data with pairwise similarity information. Subject to the scarcity of similarity information that is often expensive to collect for many application domains, existing deep learning to hash methods may overfit the training data and result in substantial loss of retrieval quality. This paper presents HashGAN, a novel architecture for deep learning to hash, which learns compact binary hash codes from both real images and diverse images synthesized by generative models. The main idea is to augment the training data with nearly real images synthesized from a new Pair Conditional Wasserstein GAN (PC-WGAN) conditioned on the pairwise similarity information. Extensive experiments demonstrate that HashGAN can generate high-quality binary hash codes and yield state-of-the-art image retrieval performance on three benchmarks, NUS-WIDE, CIFAR-10, and MS-COCO.

****

Min-Entropy Latent Model for Weakly Supervised Object Detection

Fang Wan, Pengxu Wei, Jianbin Jiao, Zhenjun Han, Qixiang Ye; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1297-1306

Weakly supervised object detection is a challenging task when provided with image category supervision but required to learn, at the same time, object locations and object detectors. The inconsistency between the weak supervision and learning objectives introduces randomness to object locations and ambiguity to detectors. In this paper, a min-entropy latent model (MELM) is proposed for weakly supervised object detection. Min-entropy is used as a metric to measure the randomness of object localization during learning, as well as serving as a model to learn object locations. It aims to principally reduce the variance of positive insta

nces and alleviate the ambiguity of detectors. MELM is deployed as two sub-models, which respectively discovers and localizes objects by minimizing the global and local entropy. MELM is unified with feature learning and optimized with a recurrent learning algorithm, which progressively transfers the weak supervision to object locations. Experiments demonstrate that MELM significantly improves the performance of weakly supervised detection, weakly supervised localization, and image classification, against the state-of-the-art approaches.
****

MAttNet: Modular Attention Network for Referring Expression Comprehension
Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, Tamara L. Berg; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1307-1315

In this paper, we address referring expression comprehension: localizing an image region described by a natural language expression. While most recent work treats expressions as a single unit, we propose to decompose them into three modular components related to subject appearance, location, and relationship to other objects. This allows us to flexibly adapt to expressions containing different types of information in an end-to-end framework. In our model, which we call the Modular Attention Network (MAttNet), two types of attention are utilized: language-based attention that learns the module weights as well as the word/phrase attention that each module should focus on; and visual attention that allows the subject and relationship modules to focus on relevant image components. Module weights combine scores from all three modules dynamically to output an overall score. Experiments show that MAttNet outperforms previous state-of-the-art methods by a large margin on both bounding-box-level and pixel-level comprehension tasks.
****

AttnGAN: Fine-Grained Text to Image Generation With Attentional Generative Adversarial Networks
Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, Xiaodong He; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1316-1324

In this paper, we propose an Attentional Generative Adversarial Network (AttnGAN) that allows attention-driven, multi-stage refinement for fine-grained text-to-image generation. With a novel attentional generative network, the AttnGAN can synthesize fine-grained details at different sub-regions of the image by paying attentions to the relevant words in the natural language description. In addition, a deep attentional multimodal similarity model is proposed to compute a fine-grained image-text matching loss for training the generator. The proposed AttnGAN significantly outperforms the previous state of the art, boosting the best reported inception score by 14.14% on the CUB dataset and 170.25% on the more challenging COCO dataset. A detailed analysis is also performed by visualizing the attention layers of the AttnGAN. It for the first time shows that the layered attentional GAN is able to automatically select the condition at the word level for generating different parts of the image.
****

Adversarial Complementary Learning for Weakly Supervised Object Localization
Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, Thomas S. Huang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1325-1334

In this work, we propose Adversarial Complementary Learning (ACoL) to automatically localize integral objects of semantic interest with weak supervision. We first mathematically prove that class localization maps can be obtained by directly selecting the class-specific feature maps of the last convolutional layer, which paves a simple way to identify object regions. We then present a simple network architecture including two parallel-classifiers for object localization. Specifically, we leverage one classification branch to dynamically localize some discriminative object regions during the forward pass. Although it is usually responsive to sparse parts of the target objects, this classifier can drive the counterpart classifier to discover new and complementary object regions by erasing its discovered regions from the feature maps. With such an adversarial learning, th e two parallel-classifiers are forced to leverage complementary object regions for classification and can finally generate integral object localization together. The merits of ACoL are mainly two-fold: 1) it can be trained in an end-to-end manner; 2) dynamically erasing enables the counterpart classifier to discover complementary object regions more effectively. We demonstrate the superiority of our ACoL approach in a variety of experiments. In particular, the Top-1 localization error rate on the ILSVRC dataset is 45.14%, which is the new state-of-the-art.

****

Conditional Generative Adversarial Network for Structured Domain Adaptation

Weixiang Hong, Zhenzhen Wang, Ming Yang, Junsong Yuan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1335-1344

In recent years, deep neural nets have triumphed over many computer vision problems, including semantic segmentation, which is a critical task in emerging autonomous driving and medical image diagnostics applications. In general, training deep neural nets requires a humongous amount of labeled data, which is laborious and costly to collect and annotate. Recent advances in computer graphics shed light on utilizing photo-realistic synthetic data with computer generated annotations to train neural nets. Nevertheless, the domain mismatch between real images and synthetic ones is the major challenge against harnessing the generated data and labels. In this paper, we propose a principled way to conduct structured domain adaption for semantic segmentation, i.e., integrating GAN into the FCN framework to mitigate the gap between source and target domains. Specifically, we learn a conditional generator to transform features of synthetic images to real-image like features, and a discriminator to distinguish them. For each training batch, the conditional generator and the discriminator compete against each other so that the generator learns to produce real-image like features to fool the discriminator; afterwards, the FCN parameters are updated to accommodate the changes of GAN. In experiments, without using labels of real image data, our method significantly outperforms the baselines as well as state-of-the-art methods by 12% ~ 20% mean IoU on the Cityscapes dataset.

****

GroupCap: Group-Based Image Captioning With Structured Relevance and Diversity Constraints

Fuhai Chen, Rongrong Ji, Xiaoshuai Sun, Yongjian Wu, Jinsong Su; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1345-1353

Most image captioning models focus on one-line (single image) captioning, where the correlations like relevance and diversity among group images (e.g., within the same album or event) are simply neglected, resulting in less accurate and diverse captions. Recent works mainly consider imposing the diversity during the online inference only, which neglect the correlation among visual structures in offline training. In this paper, we propose a novel group-based image captioning scheme (termed GroupCap), which jointly models the structured relevance and diversity among visual contents of group images towards an optimal collaborative captioning. In particular, we first propose a visual tree parser (VP-Tree) to construct the structured semantic correlations within individual images. Then, the relevance and diversity among images are well modeled by exploiting the correlations among their tree structures. Finally, such correlations are modeled as constraints and sent into the LSTM-based captioning generator. In offline optimization, we adopt an end-to-end formulation, which jointly trains the visual tree parser, the structured relevance and diversity constraints, as well as the LSTM based captioning model. To facilitate quantitative evaluation, we further release two group captioning datasets derived from the MS-COCO benchmark, serving as the first of their kind. Quantitative results show that the proposed GroupCap model outperforms the state-of-the-art and alternative approaches, which can generate much more accurate and discriminative captions under various evaluation metrics.

****

Weakly-Supervised Semantic Segmentation by Iteratively Mining Common Object Features

Xiang Wang, Shaodi You, Xi Li, Huimin Ma; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1354-1362

Weakly-supervised semantic segmentation under image tags supervision is a challenging task as it directly associates high-level semantic to low-level appearance. To bridge this gap, in this paper, we propose an iterative bottom-up and top-down framework which alternatively expands object regions and optimizes segmentation network. We start from initial localization produced by classification networks. While classification networks are only responsive to small and coarse discriminative object regions, we argue that, these regions contain significant common features about objects. So in the bottom-up step, we mine common object features from the initial localization and expand object regions with the mined features. To supplement non-discriminative regions, saliency maps are then considered under Bayesian framework to refine the object regions. Then in the top-down step, the refined object regions are used as supervision to train the segmentation network and to predict object masks. These object masks provide more accurate localization and contain more regions of object. Further, we take these object masks as initial localization and mine common object features from them. These processes are conducted iteratively to progressively produce fine object masks and optimize segmentation networks. Experimental results on Pascal VOC 2012 dataset demonstrate that the proposed method outperforms previous state-of-the-art methods by a large margin.

****

Bootstrapping the Performance of Webly Supervised Semantic Segmentation

Tong Shen, Guosheng Lin, Chunhua Shen, Ian Reid; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1363-1371

Fully supervised methods for semantic segmentation require pixel-level class masks to train, the creation of which are expensive in terms of manual labour and time. In this work, we focus on weak supervision, developing a method for training a high-quality pixel-level classifier for semantic segmentation, using only image-level class labels as the provided ground-truth. Our method is formulated as a two-stage approach in which we first aim to create accurate pixel-level masks for the training images via a bootstrapping process, and then use these now-accurately segmented images as a proxy ground-truth in a more standard supervised setting. The key driver for our work is that in the target dataset we typically have reliable ground-truth image-level labels, while data crawled from the web may have unreliable labels, but can be filtered to comprise only easy images to segment, therefore having reliable boundaries. These two forms of information are complementary and we use this observation to build a novel bi-directional transfer learning. This framework transfers knowledge between two domains, target domain and web domain, bootstrapping the performance of weakly supervised semantic segmentation. Conducting experiments on the popular benchmark dataset PASCAL VOC 2012 based on both a VGG16 network and on ResNet50, we reach state-of-the-art performance with scores of 60.2% IoU and 63.9% IoU respectively.

****

DeepVoting: A Robust and Explainable Deep Network for Semantic Part Detection Under Partial Occlusion

Zhishuai Zhang, Cihang Xie, Jianyu Wang, Lingxi Xie, Alan L. Yuille; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1372-1380

In this paper, we study the task of detecting semantic parts of an object, e.g., a wheel of a car, under partial occlusion. We propose that all models should be trained without seeing occlusions while being able to transfer the learned knowledge to deal with occlusions. This setting alleviates the difficulty in collecting an exponentially large dataset to cover occlusion patterns and is more essential. In this scenario, the proposal-based deep networks, like RCNN-series, often produce unsatisfactory results, because both the proposal extraction and classification stages may be confused by the irrelevant occluders. To address this, [25] proposed a voting mechanism that combines multiple local visual cues to detect semantic parts. The semantic parts can still be detected even though some visual cues are missing due to occlusions. However, this method is manually-designe d, thus is hard to be optimized in an end-to-end manner. In this paper, we present DeepVoting, which incorporates the robustness shown by [25] into a deep network, so that the whole pipeline can be jointly optimized. Specifically, it adds two layers after the intermediate features of a deep network, e.g., the pool-4 layer of VGGNet. The first layer extracts the evidence of local visual cues, and the second layer performs a voting mechanism by utilizing the spatial relationship between visual cues and semantic parts. We also propose an improved version DeepVoting+ by learning visual cues from context outside objects. In experiments, DeepVoting achieves significantly better performance than several baseline methods, including Faster-RCNN, for semantic part detection under occlusion. In addition, DeepVoting enjoys explainability as the detection results can be diagnosed via looking up the voting cues.

****

Geometry-Aware Scene Text Detection With Instance Transformation Network

Fangfang Wang, Liming Zhao, Xi Li, Xinchao Wang, Dacheng Tao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1381-1389

Localizing text in the wild is challenging in the situations of complicated geometric layout of the targets like random orientation and large aspect ratio. In this paper, we propose a geometry-aware modeling approach tailored for scene text representation with an end-to-end learning scheme. In our approach, a novel Instance Transformation Network (ITN) is presented to learn the geometry-aware representation encoding the unique geometric configurations of scene text instances with in-network transformation embedding, resulting in a robust and elegant framework to detect words or text lines at one pass. An end-to-end multi-task learning strategy with transformation regression, text/non-text classification and coordinate regression is adopted in the ITN. Experiments on the benchmark datasets demonstrate the effectiveness of the proposed approach in detecting scene text in various geometric configurations.

****

Optical Flow Guided Feature: A Fast and Robust Motion Representation for Video Action Recognition

Shuyang Sun, Zhanghui Kuang, Lu Sheng, Wanli Ouyang, Wei Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1390-1399

Motion representation plays a vital role in human action recognition in videos. In this study, we introduce a novel compact motion representation for video action recognition, named Optical Flow guided Feature (OFF), which enables the network to distill temporal information through a fast and robust approach. The OFF is derived from the definition of optical flow and is orthogonal to the optical flow. The derivation also provides theoretical support for using the difference between two frames. By directly calculating pixel-wise spatio-temporal gradients of the deep feature maps, the OFF could be embedded in any existing CNN based video action recognition framework with only a slight additional cost. It enables the CNN to extract spatio-temporal information, especially the temporal information between frames simultaneously. This simple but powerful idea is validated by experimental results. The network with OFF fed only by RGB inputs achieves a competitive accuracy of 93.3% on UCF-101, which is comparable with the result obtained by two streams (RGB and optical flow), but is 15 times faster in speed. Experimental results also show that OFF is complementary to other motion modalities such as optical flow. When the proposed method is plugged into the state-of-the-art video action recognition framework, it has 96.0% and 74.2% accuracy on UCF-101 and HMDB-51 respectively. The code for this project is available at: https://github.com/kevin-ssy/Optical-Flow-Guided-Feature

****

Motion-Guided Cascaded Refinement Network for Video Object Segmentation

Ping Hu, Gang Wang, Xiangfei Kong, Jason Kuen, Yap-Peng Tan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1400-1409

Deep CNNs have achieved superior performance in many tasks of computer vision an d image understanding. However, it is still difficult to effectively apply deep CNNs to video object segmentation(VOS) since treating video frames as separate and static will lose the information hidden in motion. To tackle this problem, we propose a Motion-guided Cascaded Refinement Network for VOS. By assuming the object motion is normally different from the background motion, for a video frame we first apply an active contour model on optical flow to coarsely segment objects of interest. Then, the proposed Cascaded Refinement Network(CRN) takes the coarse segmentation as guidance to generate an accurate segmentation of full resolution. In this way, the motion information and the deep CNNs can well complement each other to accurately segment objects from video frames. Furthermore, in CRN we introduce a Single-channel Residual Attention Module to incorporate the coarse segmentation map as attention, making our network effective and efficient in both training and testing. We perform experiments on the popular benchmarks and the results show that our method achieves state-of-the-art performance at a much faster speed.

****

A Memory Network Approach for Story-Based Temporal Summarization of 360° Videos

Sangho Lee, Jinyoung Sung, Youngjae Yu, Gunhee Kim; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1410-1419

We address the problem of story-based temporal summarization of long 360° videos. We propose a novel memory network model named Past-Future Memory Network (PFMN), in which we first compute the scores of 81 normal field of view (NFOV) region proposals cropped from the input 360° video, and then recover a latent, collective summary using the network with two external memories that store the embeddings of previously selected subshots and future candidate subshots. Our major contributions are two-fold. First, our work is the first to address story-based temporal summarization of 360° videos. Second, our model is the first attempt to leverage memory networks for video summarization tasks. For evaluation, we perform three sets of experiments. First, we investigate the view selection capability of our model on the Pano2Vid dataset. Second, we evaluate the temporal summarization with a newly collected 360° video dataset. Finally, we experiment our model's performance in another domain, with image-based storytelling VIST dataset. We verify that our model achieves state-of-the-art performance on all the tasks.

****

Cube Padding for Weakly-Supervised Saliency Prediction in 360° Videos

Hsien-Tzu Cheng, Chun-Hung Chao, Jin-Dong Dong, Hao-Kai Wen, Tyng-Luh Liu, Min Sun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1420-1429

Automatic saliency prediction in 360° videos is critical for viewpoint guidance applications (e.g., Facebook 360 Guide). We propose a spatial-temporal network which is (1) unsupervisedly trained and (2) tailor-made for 360° viewing sphere. Note that most existing methods are less scalable since they rely on annotated saliency map for training. Most importantly, they convert 360° sphere to 2D images (e.g., a single equirectangular image or multiple separate Normal Field-of-View (NFoV) images) which introduces distortion and image boundaries. In contrast, we propose a simple and effective Cube Padding (CP) technique as follows. Firstly, we render the 360° view on six faces of a cube using perspective projection. Thus, it introduces very little distortion. Then, we concatenate all six faces while utilizing the connectivity between faces on the cube for image padding (i.e., Cube Padding) in convolution, pooling, convolutional LSTM layers. In this way, PC introduces no image boundary while being applicable to almost all Convolutional Neural Network (CNN) structures. To evaluate our method, we propose Wild-360, a new 360° video saliency dataset, containing challenging videos with saliency heatmap annotations. In experiments, our method outperforms all baseline methods in both speed and quality.

****

Appearance-and-Relation Networks for Video Classification

Limin Wang, Wei Li, Wen Li, Luc Van Gool; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1430-1439

Spatiotemporal feature learning in videos is a fundamental problem in computer v

ision. This paper presents a new architecture, termed as Appearance-and-Relation Network (ARTNet), to learn video representation in an end-to-end manner. ARTNets are constructed by stacking multiple generic building blocks, called as SMART, whose goal is to simultaneously model appearance and relation from RGB input in a separate and explicit manner. Specifically, SMART blocks decouple the spatiotemporal learning module into an appearance branch for spatial modeling and a relation branch for temporal modeling. The appearance branch is implemented based on the linear combination of pixels or filter responses in each frame, while the relation branch is designed based on the multiplicative interactions between pixels or filter responses across multiple frames. We perform experiments on three action recognition benchmarks: Kinetics, UCF101, and HMDB51, demonstrating that SMART blocks obtain an evident improvement over 3D convolutions for spatiotemporal feature learning. Under the same training setting, ARTNets achieve superior performance on these three datasets to the existing state-of-the-art methods.
****

Excitation Backprop for RNNs
Sarah Adel Bargal, Andrea Zunino, Donghyun Kim, Jianming Zhang, Vittorio Murino, Stan Sclaroff; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1440-1449

Deep models are state-of-the-art or many vision tasks including video action recognition and video captioning. Models are trained to caption or classify activity in videos, but little is known about the evidence used to make such decisions. Grounding decisions made by deep networks has been studied in spatial visual content, giving more insight into model predictions for images. However, such studies are relatively lacking for models of spatiotemporal visual content - videos. In this work, we devise a formulation that simultaneously grounds evidence in space and time, in a single pass, using top-down saliency. We visualize the spatiotemporal cues that contribute to a deep model's classification/captioning output using the model's internal representation. Based on these spatiotemporal cues, we are able to localize segments within a video that correspond with a specific action, or phrase from a caption, without explicitly optimizing/training for these tasks.
****

One-Shot Action Localization by Learning Sequence Matching Network
Hongtao Yang, Xuming He, Fatih Porikli; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1450-1459

Learning based temporal action localization methods require vast amounts of training data. However, such large-scale video datasets, which are expected to capture the dynamics of every action category, are not only very expensive to acquire but are also not practical simply because there exists an uncountable number of action classes. This poses a critical restriction to the current methods when the training samples are few and rare (e.g. when the target action classes are not present in the current publicly available datasets). To address this challenge, we conceptualize a new example-based action detection problem where only a few examples are provided, and the goal is to find the occurrences of these examples in an untrimmed video sequence. Towards this objective, we introduce a novel one-shot action localization method that alleviates the need for large amounts of training samples. Our solution adopts the one-shot learning technique of Matching Network and utilizes correlations to mine and localize actions of previously unseen classes. We evaluate our one-shot action localization method on the THUMOS14 and ActivityNet datasets, of which we modified the configuration to fit our one-shot problem setup.
****

Structure Preserving Video Prediction
Jingwei Xu, Bingbing Ni, Zefan Li, Shuo Cheng, Xiaokang Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1460-1469

Despite recent emergence of adversarial based methods for video prediction, existing algorithms often produce unsatisfied results in image regions with rich structural information (i.e., object boundary) and detailed motion (i.e., articulat ed body movement). To this end, we present a structure preserving video prediction framework to explicitly address above issues and enhance video prediction quality. On one hand, our framework contains a two-stream generation architecture which deals with high frequency video content (i.e., detailed object or articulated motion structure) and low frequency video content (i.e., location or moving directions) in two separate streams. On the other hand, we propose a RNN structure for video prediction, which employs temporal-adaptive convolutional kernels to capture time-varying motion patterns as well as the tiny object within a scene. Extensive experiments on diverse scene, ranging from human motion to semantic layout prediction, demonstrate the effectiveness of the proposed video prediction approach.

****

Person Re-Identification With Cascaded Pairwise Convolutions

Yicheng Wang, Zhenzhong Chen, Feng Wu, Gang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1470-1478

In this paper, a novel deep architecture named BraidNet is proposed for person re-identification. BraidNet has a specially designed WConv layer, and the cascaded WConv structure learns to extract the comparison features of two images, which are robust to misalignments and color differences across cameras. Furthermore, a Channel Scaling layer is designed to optimize the scaling factor of each input channel, which helps mitigate the zero gradient problem in the training phase. To solve the problem of imbalanced volume of negative and positive training samples, a Sample Rate Learning strategy is proposed to adaptively update the ratio between positive and negative samples in each batch. Experiments conducted on CUHK03-Detected, CUHK03-Labeled, CUHK01, Market-1501 and DukeMTMC-reID datasets demonstrate that our method achieves competitive performance when compared to state-of-the-art methods.

****

On the Importance of Label Quality for Semantic Segmentation

Aleksandar Zlateski, Ronnachai Jaroensri, Prafull Sharma, Frédo Durand; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1479-1487

Convolutional networks (ConvNets) have become the dominant approach to semantic image segmentation. Producing accurate, pixel--level labels required for this task is a tedious and time consuming process; however, producing approximate, coarse labels could take only a fraction of the time and effort. We investigate the relationship between the quality of labels and the performance of ConvNets for semantic segmentation. We create a very large synthetic dataset with perfectly labeled street view scenes. From these perfect labels, we synthetically coarsen labels with different qualities and estimate human--hours required for producing them. We perform a series of experiments by training ConvNets with a varying number of training images and label quality. We found that the performance of ConvNets mostly depends on the time spent creating the training labels. That is, a larger coarsely--annotated dataset can yield the same performance as a smaller finely--annotated one. Furthermore, fine--tuning coarsely pre--trained ConvNets with few finely-annotated labels can yield comparable or superior performance to training it with a large amount of finely-annotated labels alone, at a fraction of the labeling cost. We demonstrate that our result is also valid for different network architectures, and various object classes in an urban scene.

****

Scalable and Effective Deep CCA via Soft Decorrelation

Xiaobin Chang, Tao Xiang, Timothy M. Hospedales; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1488-1497

Recently the widely used multi-view learning model, Canonical Correlation Analysis (CCA) has been generalised to the non-linear setting via deep neural networks. Existing deep CCA models typically first decorrelate the feature dimensions of each view before the different views are maximally correlated in a common latent space. This feature decorrelation is achieved by enforcing an exact decorrelation constraint; these models are thus computationally expensive due to the matrix inversion or SVD operations required for exact decorrelation at each training

iteration. Furthermore, the decorrelation step is often separated from the gradient descent based optimisation, resulting in sub-optimal solutions. We propose a novel deep CCA model Soft CCA to overcome these problems. Specifically, exact decorrelation is replaced by soft decorrelation via a mini-batch based Stochastic Decorrelation Loss (SDL) to be optimised jointly with the other training objectives. Extensive experiments show that the proposed soft CCA is more effective and efficient than existing deep CCA models. In addition, our SDL loss can be applied to other deep models beyond multi-view learning, and obtains superior performance compared to existing decorrelation losses.
****

Duplex Generative Adversarial Network for Unsupervised Domain Adaptation
Lanqing Hu, Meina Kan, Shiguang Shan, Xilin Chen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1498-1507
Domain adaptation attempts to transfer the knowledge obtained from the source domain to the target domain, i.e., the domain where the testing data are. The main challenge lies in the distribution discrepancy between source and target domain. Most existing works endeavor to learn domain invariant representation usually by minimizing a distribution distance, e.g., MMD and the discriminator in the recently proposed generative adversarial network (GAN). Following the similar idea of GAN, this work proposes a novel GAN architecture with duplex adversarial discriminators (referred to as DupGAN), which can achieve domain-invariant representation and domain transformation. Specifically, our proposed network consists of three parts, an encoder, a generator and two discriminators. The encoder embeds samples from both domains into the latent representation, and the generator decodes the latent representation to both source and target domains respectively conditioned on a domain code, i.e., achieves domain transformation. The generator is pitted against duplex discriminators, one for source domain and the other for target, to ensure the reality of domain transformation, the latent representation domain invariant and the category information of it preserved as well. Our proposed work achieves the state-of-the-art performance on unsupervised domain adaptation of digit classification and object recognition.
****

Edit Probability for Scene Text Recognition
Fan Bai, Zhanzhan Cheng, Yi Niu, Shiliang Pu, Shuigeng Zhou; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1508-1516
We consider the scene text recognition problem under the attention-based encoder-decoder framework, which is the state of the art. The existing methods usually employ a frame-wise maximal likelihood loss to optimize the models. When we train the model, the misalignment between the ground truth strings and the attention's output sequences of probability distribution, which is caused by missing or superfluous characters, will confuse and mislead the training process, and consequently make the training costly and degrade the recognition accuracy. To handle this problem, we propose a novel method called edit probability (EP) for scene text recognition. EP tries to effectively estimate the probability of generating a string from the output sequence of probability distribution conditioned on the input image, while considering the possible occurrences of missing/superfluous characters. The advantage lies in that the training process can focus on the missing, superfluous and unrecognized characters, and thus the impact of the misalignment problem can be alleviated or even overcome. We conduct extensive experiments on standard benchmarks, including the IIIT-5K, Street View Text and ICDAR datasets. Experimental results show that the EP can substantially boost scene text recognition performance.
****

Global Versus Localized Generative Adversarial Nets
Guo-Jun Qi, Liheng Zhang, Hao Hu, Marzieh Edraki, Jingdong Wang, Xian-Sheng Hua; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1517-1525
In this paper, we present a novel localized Generative Adversarial Net (GAN) to learn on the manifold of real data. Compared with the classic GAN that {em globa

lly} parameterizes a manifold, the Localized GAN (LGAN) uses local coordinate charts to parameterize distinct local geometry of how data points can transform at different locations on the manifold. Specifically, around each point there exists a {em local} generator that can produce data following diverse patterns of transformations on the manifold. The locality nature of LGAN enables local generators to adapt to and directly access the local geometry without need to invert the generator in a global GAN. Furthermore, it can prevent the manifold from being locally collapsed to a dimensionally deficient tangent subspace by imposing an orthonormality prior between tangents. This provides a geometric approach to alleviating mode collapse at least locally on the manifold by imposing independence between data transformations in different tangent directions. We will also demonstrate the LGAN can be applied to train a robust classifier that prefers locally consistent classification decisions on the manifold, and the resultant regularizer is closely related with the Laplace-Beltrami operator. Our experiments show that the proposed LGANs can not only produce diverse image transformations, but also deliver superior classification performances.

****

MoCoGAN: Decomposing Motion and Content for Video Generation

Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, Jan Kautz; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1526-1535

Visual signals in a video can be divided into content and motion. While content specifies which objects are in the video, motion describes their dynamics. Based on this prior, we propose the Motion and Content decomposed Generative Adversarial Network (MoCoGAN) framework for video generation. The proposed framework generates a video by mapping a sequence of random vectors to a sequence of video frames. Each random vector consists of a content part and a motion part. While the content part is kept fixed, the motion part is realized as a stochastic process. To learn motion and content decomposition in an unsupervised manner, we introduce a novel adversarial learning scheme utilizing both image and video discriminators. Extensive experimental results on several challenging datasets with qualitative and quantitative comparison to the state-of-the-art approaches, verify effectiveness of the proposed framework. In addition, we show that MoCoGAN allows one to generate videos with same content but different motion as well as videos with different content and same motion. Our code is available at https://github.com/sergeytulyakov/mocogan.

****

Recurrent Residual Module for Fast Inference in Videos

Bowen Pan, Wuwei Lin, Xiaolin Fang, Chaoqin Huang, Bolei Zhou, Cewu Lu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1536-1545

Deep convolutional neural networks (CNNs) have made impressive progress in many video recognition tasks such as video pose estimation and video object detection. However, running CNN inference on video requires numerous computation and is usually slow. In this work, we propose a framework called Recurrent Residual Module (RRM) to accelerate the CNN inference for video recognition tasks. This framework has a novel design of using the similarity of the intermediate feature maps of two consecutive frames to largely reduce the redundant computation. One unique property of the proposed method compared to previous work is that feature maps of each frame are precisely computed. The experiments show that, while maintaining the similar recognition performance, our RRM yields averagely 2× acceleration on the commonly used CNNs such as AlexNet, ResNet, deep compression model (thus 8–12× faster than the original dense models on the efficient inference engine), and impressively 9× acceleration on some binary networks such as XNOR-Nets (thus 500× faster than the original model). We further verify the effectiveness of the RRM on speeding CNNs for video pose estimation and video object detection.

****

Improving Landmark Localization With Semi-Supervised Learning

Sina Honari, Pavlo Molchanov, Stephen Tyree, Pascal Vincent, Christopher Pal, Jan Kautz; Proceedings of the IEEE Conference on Computer Vision and Pattern Recog nition (CVPR), 2018, pp. 1546-1555

We present two techniques to improve landmark localization in images from partially annotated datasets. Our primary goal is to leverage the common situation where precise landmark locations are only provided for a small data subset, but where class labels for classification or regression tasks related to the landmarks are more abundantly available. First, we propose the framework of sequential multitasking and explore it here through an architecture for landmark localization where training with class labels acts as an auxiliary signal to guide the landmark localization on unlabeled data. A key aspect of our approach is that errors can be backpropagated through a complete landmark localization model. Second, we propose and explore an unsupervised learning technique for landmark localization based on having a model predict equivariant landmarks with respect to transformations applied to the image. We show that these techniques, improve landmark prediction considerably and can learn effective detectors even when only a small fraction of the dataset has landmark labels. We present results on two toy datasets and four real datasets, with hands and faces, and report new state-of-the-art on two datasets in the wild, e.g. with only 5% of labeled images we outperform previous state-of-the-art trained on the AFLW dataset.

****

Adversarial Data Programming: Using GANs to Relax the Bottleneck of Curated Labeled Data

Arghya Pal, Vineeth N. Balasubramanian; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1556-1565

Paucity of large curated hand labeled training data forms a major bottleneck in the deployment of machine learning models in computer vision and other fields. Recent work (Data Programming) has shown how distant supervision signals in the form of labeling functions can be used to obtain labels for given data in near-constant time. In this work, we present Adversarial Data Programming (ADP), which presents an adversarial methodology to generate data as well as a curated aggregated label, given a set of weak labeling functions. We validated our method on the MNIST, Fashion MNIST, CIFAR 10 and SVHN datasets, and it outperformed many state-of-the-art models. We conducted extensive experiments to study its usefulness, as well as showed how the proposed ADP framework can be used for transfer learning as well as multitask learning, where data from two domains are generated simultaneously using the framework along with the label information. Our future work will involve understanding the theoretical implications of this new framework from a game-theoretic perspective, as well as explore the performance of the method on more complex datasets.

****

Stochastic Variational Inference With Gradient Linearization

Tobias Plötz, Anne S. Wannenwetsch, Stefan Roth; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1566-1575

Variational inference has experienced a recent surge in popularity owing to stochastic approaches, which have yielded practical tools for a wide range of model classes. A key benefit is that stochastic variational inference obviates the tedious process of deriving analytical expressions for closed-form variable updates. Instead, one simply needs to derive the gradient of the log-posterior, which is often much easier. Yet for certain model classes, the log-posterior itself is difficult to optimize using standard gradient techniques. One such example are random field models, where optimization based on gradient linearization has proven popular, since it speeds up convergence significantly and can avoid poor local optima. In this paper we propose stochastic variational inference with gradient linearization (SVIGL). It is similarly convenient as standard stochastic variational inference - all that is required is a local linearization of the energy gradient. Its benefit over stochastic variational inference with conventional gradient methods is a clear improvement in convergence speed, while yielding comparable or even better variational approximations in terms of KL divergence. We demonstrate the benefits of SVIGL in three applications: Optical flow estimation, Poisson-Gaussian denoising, and 3D surface reconstruction.

****

Multi-Label Zero-Shot Learning With Structured Knowledge Graphs
Chung-Wei Lee, Wei Fang, Chih-Kuan Yeh, Yu-Chiang Frank Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1576-1585

In this paper, we propose a novel deep learning architecture for multi-label zero-shot learning (ML-ZSL), which is able to predict multiple unseen class labels for each input instance. Inspired by the way humans utilize semantic knowledge between objects of interests, we propose a framework that incorporates knowledge graphs for describing the relationships between multiple labels. Our model learns an information propagation mechanism from the semantic label space, which can be applied to model the interdependencies between seen and unseen class labels. With such investigation of structured knowledge graphs for visual reasoning, we show that our model can be applied for solving multi-label classification and ML-ZSL tasks. Compared to state-of-the-art approaches, comparable or improved performances can be achieved by our method.

****

MorphNet: Fast & Simple Resource-Constrained Structure Learning of Deep Networks
Ariel Gordon, Elad Eban, Ofir Nachum, Bo Chen, Hao Wu, Tien-Ju Yang, Edward Choi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1586-1595

We present MorphNet, an approach to automate the design of neural network structures. MorphNet iteratively shrinks and expands a network, shrinking via a resource-weighted sparsifying regularizer on activations and expanding via a uniform multiplicative factor on all layers. In contrast to previous approaches, our method is scalable to large networks, adaptable to specific resource constraints (e.g. the number of floating-point operations per inference), and capable of increasing the network’s performance. When applied to standard network architectures on a wide variety of datasets, our approach discovers novel structures in each domain, obtaining higher performance while respecting the resource constraint.

****

Deep Adversarial Subspace Clustering
Pan Zhou, Yunqing Hou, Jiashi Feng; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1596-1604

Most existing subspace clustering methods hinge on self-expression of handcrafted representations and are unaware of potential clustering errors. Thus they perform unsatisfactorily on real data with complex underlying subspaces. To solve this issue, we propose a novel deep adversarial subspace clustering (DASC) model, which learns more favorable sample representations by deep learning for subspace clustering, and more importantly introduces adversarial learning to supervise sample representation learning and subspace clustering. Specifically, DASC consists of a subspace clustering generator and a quality-verifying discriminator, which learn against each other. The generator produces subspace estimation and sample clustering. The discriminator evaluates current clustering performance by inspecting whether the re-sampled data from estimated subspaces have consistent subspace properties, and supervises the generator to progressively improve subspace clustering. Experimental results on the handwritten recognition, face and object clustering tasks demonstrate the advantages of DASC over shallow and few deep subspace clustering models. Moreover, to our best knowledge, this is the first successful application of GAN-alike model for unsupervised subspace clustering, which also paves the way for deep learning to solve other unsupervised learning problems.

****

Towards Human-Machine Cooperation: Self-Supervised Sample Mining for Object Detection
Keze Wang, Xiaopeng Yan, Dongyu Zhang, Lei Zhang, Liang Lin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1605-1613

Though quite challenging, leveraging large-scale unlabeled or partially labeled images in a cost-effective way has increasingly attracted interests for its grea t importance to computer vision. To tackle this problem, many Active Learning (AL) methods have been developed. However, these methods mainly define their sample selection criteria within a single image context, leading to the suboptimal robustness and impractical solution for large-scale object detection. In this paper, aiming to remedy the drawbacks of existing AL methods, we present a principled Self-supervised Sample Mining (SSM) process accounting for the real challenges in object detection. Specifically, our SSM process concentrates on automatically discovering and pseudo-labeling reliable region proposals for enhancing the object detector via the introduced cross image validation, i.e., pasting these proposals into different labeled images to comprehensively measure their values under different image contexts. By resorting to the SSM process, we propose a new AL framework for gradually incorporating unlabeled or partially labeled data into the model learning while minimizing the annotating effort of users. Extensive experiments on two public benchmarks clearly demonstrate our proposed framework can achieve the comparable performance to the state-of-the-art methods with significantly fewer annotations.

****

Discrete-Continuous ADMM for Transductive Inference in Higher-Order MRFs

Emanuel Laude, Jan-Hendrik Lange, Jonas Schüpfer, Csaba Domokos, Laura Leal-Taixé, Frank R. Schmidt, Bjoern Andres, Daniel Cremers; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1614-1624

This paper introduces a novel algorithm for transductive inference in higher-order MRFs, where the unary energies are parameterized by a variable classifier. The considered task is posed as a joint optimization problem in the continuous classifier parameters and the discrete label variables. In contrast to prior approaches such as convex relaxations, we propose an advantageous decoupling of the objective function into discrete and continuous subproblems and a novel, efficient optimization method related to ADMM. This approach preserves integrality of the discrete label variables and guarantees global convergence to a critical point. We demonstrate the advantages of our approach in several experiments including video object segmentation on the DAVIS data set and interactive image segmentation.

****

Robust Physical-World Attacks on Deep Learning Visual Classification

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, Dawn Song; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1625-1634

Recent studies show that the state-of-the-art deep neural networks (DNNs) are vulnerable to adversarial examples, resulting from small-magnitude perturbations added to the input. Given that that emerging physical systems are using DNNs in safety-critical situations, adversarial examples could mislead these systems and cause dangerous situations. Therefore, understanding adversarial examples in the physical world is an important step towards developing resilient learning algorithms. We propose a general attack algorithm, Robust Physical Perturbations (RP 2), to generate robust visual adversarial perturbations under different physical conditions. Using the real-world case of road sign classification, we show that adversarial examples generated using RP 2 achieve high targeted misclassification rates against standard-architecture road sign classifiers in the physical world under various environmental conditions, including viewpoints. Due to the current lack of a standardized testing method, we propose a two-stage evaluation methodology for robust physical adversarial examples consisting of lab and field tests. Using this methodology, we evaluate the efficacy of physical adversarial manipulations on real objects. With a perturbation in the form of only black and white stickers, we attack a real stop sign, causing targeted misclassification in 100% of the images obtained in lab settings, and in 84.8% of the captured video frames obtained on a moving vehicle (field test) for the target classifier.

****

Generating a Fusion Image: One's Identity and Another's Shape

DongGyu Joo, Doyeon Kim, Junmo Kim; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1635-1643

Generating a novel image by manipulating two input images is an interesting research problem in the study of generative adversarial networks (GANs). We propose a new GAN-based network that generates a fusion image with the identity of input image x and the shape of input image y. Our network can simultaneously train on more than two image datasets in an unsupervised manner. We define an identity loss LI to catch the identity of image x and a shape loss LS to get the shape of y. In addition, we propose a novel training method called Min-Patch training to focus the generator on crucial parts of an image, rather than its entirety. We show qualitative results on the VGG Youtube Pose dataset , Eye dataset (MPIIGaze and UnityEyes), and the Photo–Sketch–Cartoon dataset.

****

Learning to Promote Saliency Detectors

Yu Zeng, Huchuan Lu, Lihe Zhang, Mengyang Feng, Ali Borji; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1644-1653

The categories and appearance of salient objects vary from image to image, therefore, saliency detection is an image-specific task. Due to lack of large-scale saliency training data, using deep neural networks (DNNs) with pre-training is difficult to precisely capture the image-specific saliency cues. To solve this issue, we formulate a zero-shot learning problem to promote existing saliency detectors. Concretely, a DNN is trained as an embedding function to map pixels and the attributes of the salient/background regions of an image into the same metric space, in which an image-specific classifier is learned to classify the pixels. Since the image-specific task is performed by the classifier, the DNN embedding effectively plays the role of a general feature extractor. Compared with transferring the learning to a new recognition task using limited data, this formulation makes the DNN learn more effectively from small data. Extensive experiments on five data sets show that our method significantly improves accuracy of existing methods and compares favorably against state-of-the-art approaches.

****

Image Super-Resolution via Dual-State Recurrent Networks

Wei Han, Shiyu Chang, Ding Liu, Mo Yu, Michael Witbrock, Thomas S. Huang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1654-1663

Advances in image super-resolution (SR) have recently benefited significantly from rapid developments in deep neural networks. Inspired by these recent discoveries, we note that many state-of-the-art deep SR architectures can be reformulated as a single-state recurrent neural network (RNN) with finite unfoldings. In this paper, we explore new structures for SR based on this compact RNN view, leading us to a dual-state design, the Dual-State Recurrent Network (DSRN). Compared to its single-state counterparts that op- erate at a fixed spatial resolution, DSRN exploits both low- resolution (LR) and high-resolution (HR) signals jointly. Recurrent signals are exchanged between these states in both directions (both LR to HR and HR to LR) via de- layed feedback. Extensive quantitative and qualitative eval- uations on benchmark datasets and on a recent challenge demonstrate that the proposed DSRN performs favorably against state-of-the-art algorithms in terms of both mem- ory consumption and predictive accuracy.

****

Deep Back-Projection Networks for Super-Resolution

Muhammad Haris, Gregory Shakhnarovich, Norimichi Ukita; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1664-1673

The feed-forward architectures of recently proposed deep super-resolution networks learn representations of low-resolution inputs, and the non-linear mapping from those to high-resolution output. However, this approach does not fully address the mutual dependencies of low- and high-resolution images. We propose Deep Back-Projection Networks (DBPN), that exploit iterative up- and down-sampling layers, providing an error feedback mechanism for projection errors at each stage. We construct mutually-connected up- and down-sampling stages each of which represents different types of image degradation and high-resolution components. We sho w that extending this idea to allow concatenation of features across up- and down-sampling stages (Dense DBPN) allows us to reconstruct further improve super-resolution, yielding superior results and in particular establishing new state of the art results for large scaling factors such as 8x across multiple data sets.

****

Focus Manipulation Detection via Photometric Histogram Analysis
Can Chen, Scott McCloskey, Jingyi Yu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1674-1682

With the rise of misinformation spread via social media channels, enabled by the increasing automation and realism of image manipulation tools, image forensics is an increasingly relevant problem. Classic image forensic methods leverage low-level cues such as metadata, sensor noise fingerprints, and others that are easily fooled when the image is re-encoded upon upload to facebook, etc. This necessitates the use of higher-level physical and semantic cues that, once hard to estimate reliably in the wild, have become more effective due to the increasing power of computer vision. In particular, we detect manipulations introduced by artificial blurring of the image, which creates inconsistent photometric relationships between image intensity and various cues. We achieve 98% accuracy on the most challenging cases in a new dataset of blur manipulations, where the blur is geometrically correct and consistent with the scene's physical arrangement. Such manipulations are now easily generated, for instance, by smartphone cameras having hardware to measure depth, e.g. `Portrait Mode' of the iPhone7Plus. We also demonstrate good performance on a challenge dataset evaluating a wider range of manipulations in imagery representing `in the wild' conditions.

****

Compassionately Conservative Balanced Cuts for Image Segmentation
Nathan D. Cahill, Tyler L. Hayes, Renee T. Meinhold, John F. Hamilton; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1683-1691

The Normalized Cut (NCut) objective function, widely used in data clustering and image segmentation, quantifies the cost of graph partitioning in a way that biases clusters or segments that are balanced towards having lower values than unbalanced partitionings. However, this bias is so strong that it avoids any singleton partitions, even when vertices are very weakly connected to the rest of the graph. Motivated by the Buehler-Hein family of balanced cut costs, we propose the family of Compassionately Conservative Balanced (CCB) Cut costs, which are indexed by a parameter that can be used to strike a compromise between the desire to avoid too many singleton partitions and the notion that all partitions should be balanced. We show that CCB-Cut minimization can be relaxed into an orthogonally constrained $ell_{■au}$-minimization problem that coincides with the problem of computing Piecewise Flat Embeddings (PFE) for one particular index value, and we present an algorithm for solving the relaxed problem by iteratively minimizing a sequence of reweighted Rayleigh quotients (IRRQ). Using images from the BSDS500 database, we show that image segmentation based on CCB-Cut minimization provides better accuracy with respect to ground truth and greater variability in region size than NCut-based image segmentation.

****

A High-Quality Denoising Dataset for Smartphone Cameras
Abdelrahman Abdelhamed, Stephen Lin, Michael S. Brown; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1692-1700

The last decade has seen an astronomical shift from imaging with DSLR and point-and-shoot cameras to imaging with smartphone cameras. Due to the small aperture and sensor size, smartphone images have notably more noise than their DSLR counterparts. While denoising for smartphone images is an active research area, the research community currently lacks a denoising image dataset representative of real noisy images from smartphone cameras with high-quality ground truth. We address this issue in this paper with the following contributions.   We propose a systematic procedure for estimating ground truth for noisy images that can be used to benchmark denoising performance for smartphone cameras. Using this procedure, we have captured a dataset, the Smartphone Image Denoising Dataset (SIDD), of ~30,000 noisy images from 10 scenes under different lighting conditions using five representative smartphone cameras and generated their ground truth images. We used this dataset to benchmark a number of denoising algorithms. We show that CNN-based methods perform better when trained on our high-quality dataset than when trained using alternative strategies, such as low-ISO images used as a proxy for ground truth data.

****

Context-Aware Synthesis for Video Frame Interpolation

Simon Niklaus, Feng Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1701-1710

Video frame interpolation algorithms typically estimate optical flow or its variations and then use it to guide the synthesis of an intermediate frame between two consecutive original frames. To handle challenges like occlusion, bidirectional flow between the two input frames is often estimated and used to warp and blend the input frames. However, how to effectively blend the two warped frames still remains a challenging problem. This paper presents a context-aware synthesis approach that warps not only the input frames but also their pixel-wise contextual information and uses them to interpolate a high-quality intermediate frame. Specifically, we first use a pre-trained neural network to extract per-pixel contextual information for input frames. We then employ a state-of-the-art optical flow algorithm to estimate bidirectional flow between them and pre-warp both input frames and their context maps. Finally, unlike common approaches that blend the pre-warped frames, our method feeds them and their context maps to a video frame synthesis neural network to produce the interpolated frame in a context-aware fashion. Our neural network is fully convolutional and is trained end to end. Our experiments show that our method can handle challenging scenarios such as occlusion and large motion and outperforms representative state-of-the-art approaches.

****

Salient Object Detection Driven by Fixation Prediction

Wenguan Wang, Jianbing Shen, Xingping Dong, Ali Borji; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1711-1720

Research in visual saliency has been focused on two major types of models namely fixation prediction and salient object detection. The relationship between the two, however, has been less explored. In this paper, we propose to employ the former model type to identify and segment salient objects in scenes. We build a novel neural network called Attentive Saliency Network (ASNet) that learns to detect salient objects from fixation maps. The fixation map, derived at the upper network layers, captures a high-level understanding of the scene. Salient object detection is then viewed as fine-grained object-level saliency segmentation and is progressively optimized with the guidance of the fixation map in a top-down manner. ASNet is based on a hierarchy of convolutional LSTMs (convLSTMs) that offers an efficient recurrent mechanism for sequential refinement of the segmentation map. Several loss functions are introduced for boosting the performance of the ASNet. Extensive experimental evaluation shows that our proposed ASNet is capable of generating accurate segmentation maps with the help of the computed fixation map. Our work offers a deeper insight into the mechanisms of attention and narrows the gap between salient object detection and fixation prediction.

****

Enhancing the Spatial Resolution of Stereo Images Using a Parallax Prior

Daniel S. Jeon, Seung-Hwan Baek, Inchang Choi, Min H. Kim; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1721-1730

We present a novel method that can enhance the spatial resolution of stereo images using a parallax prior. While traditional stereo imaging has focused on estimating depth from stereo images, our method utilizes stereo images to enhance spatial resolution instead of estimating disparity. The critical challenge for enhancing spatial resolution from stereo images: how to register corresponding pixels with subpixel accuracy. Since disparity in traditional stereo imaging is calculated per pixel, it is directly inappropriate for enhancing spatial resolution.

We, therefore, learn a parallax prior from stereo image datasets by jointly training two-stage networks. The first network learns how to enhance the spatial resolution of stereo images in luminance, and the second network learns how to reconstruct a high-resolution color image from high-resolution luminance and chrominance of the input image. Our two-stage joint network enhances the spatial resolution of stereo images significantly more than single-image super-resolution methods. The proposed method is directly applicable to any stereo depth imaging methods, enabling us to enhance the spatial resolution of stereo images.

****

HATS: Histograms of Averaged Time Surfaces for Robust Event-Based Object Classification

Amos Sironi, Manuele Brambilla, Nicolas Bourdis, Xavier Lagorce, Ryad Benosman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1731-1740

Event-based cameras have recently drawn the attention of the Computer Vision community thanks to their advantages in terms of high temporal resolution, low power consumption and high dynamic range, compared to traditional frame-based cameras. These properties make event-based cameras an ideal choice for autonomous vehicles, robot navigation or UAV vision, among others. However, the accuracy of event-based object classification algorithms, which is of crucial importance for any reliable system working in real-world conditions, is still far behind their frame-based counterparts. Two main reasons for this performance gap are: 1. The lack of effective low-level representations and architectures for event-based object classification and 2. The absence of large real-world event-based datasets. In this paper we address both problems. First, we introduce a novel event-based feature representation together with a new machine learning architecture. Compared to previous approaches, we use local memory units to efficiently leverage past temporal information and build a robust event-based representation. Second, we release the first large real-world event-based dataset for object classification. We compare our method to the state-of-the-art with extensive experiments, showing better classification performance and real-time computation.

****

A Bi-Directional Message Passing Model for Salient Object Detection

Lu Zhang, Ju Dai, Huchuan Lu, You He, Gang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1741-1750

Recent progress on salient object detection is beneficial from Fully Convolutional Neural Network (FCN). The saliency cues contained in multi-level convolutional features are complementary for detecting salient objects. How to integrate multi-level features becomes an open problem in saliency detection. In this paper, we propose a novel bi-directional message passing model to integrate multi-level features for salient object detection. At first, we adopt a Multi-scale Context-aware Feature Extraction Module (MCFEM) for multi-level feature maps to capture rich context information. Then a bi-directional structure is designed to pass messages between multi-level features, and a gate function is exploited to control the message passing rate. We use the features after message passing, which simultaneously encode semantic information and spatial details, to predict saliency maps. Finally, the predicted results are efficiently combined to generate the final saliency map. Quantitative and qualitative experiments on five benchmark datasets demonstrate that our proposed model performs favorably against the state-of-the-art methods under different evaluation metrics.

****

Matching Pixels Using Co-Occurrence Statistics

Rotal Kat, Roy Jevnisek, Shai Avidan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1751-1759

We propose a new error measure for matching pixels that is based on co-occurrence statistics. The measure relies on a co-occurrence matrix that counts the number of times pairs of pixel values co-occur within a window. The error incurred by matching a pair of pixels is inverse proportional to the probability that their values co-occur together, and not their color difference. This measure also works with features other than color, e.g. deep features. We show that this improve s the state-of-the-art performance of template matching on standard benchmarks. We then propose an embedding scheme that maps the input image to an embedded image such that the Euclidean distance between pixel values in the embedded space resembles the co-occurrence statistics in the original space. This lets us run existing vision algorithms on the embedded images and enjoy the power of co-occurrence statistics for free. We demonstrate this on two algorithms, the Lucas-Kanade image registration and the Kernelized Correlation Filter (KCF) tracker. Experiments show that performance of each algorithm improves by about 10%.

****

SeedNet: Automatic Seed Generation With Deep Reinforcement Learning for Robust Interactive Segmentation

Gwangmo Song, Heesoo Myeong, Kyoung Mu Lee; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1760-1768

In this paper, we propose an automatic seed generation technique with deep reinforcement learning to solve the interactive segmentation problem. One of the main issues of the interactive segmentation problem is robust and consistent object extraction with less human effort. Most of the existing algorithms highly depend on the distribution of inputs, which differs from one user to another and hence need sequential user interactions to achieve adequate performance. In our system, when a user first specifies a point on the desired object and a point in the background, a sequence of artificial user input is automatically generated for precisely segmenting the desired object. The proposed system allows the user to reduce the number of input significantly. This problem is difficult to cast as a supervised learning problem because it is not possible to define globally optimal user input at some stage of the interactive segmentation task. Hence, we formulate automatic seed generation problem as Markov Decision Process (MDP) and then optimize it by reinforcement learning with Deep Q-Network (DQN). We train our network on the MSRA10K dataset and show that the network achieves notable performance improvement from inaccurate initial segmentation on both seen and unseen datasets.

****

Jerk-Aware Video Acceleration Magnification

Shoichiro Takeda, Kazuki Okami, Dan Mikami, Megumi Isogai, Hideaki Kimata; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1769-1777

Video magnification reveals subtle changes invisible to the naked eye, but such tiny yet meaningful changes are often hidden under large motions: small deformation of the muscles in doing sports, or tiny vibrations of strings in ukulele playing. For magnifying subtle changes under large motions, video acceleration magnification method has recently been proposed. This method magnifies subtle acceleration changes and ignores slow large motions. However, quick large motions severely distort this method. In this paper, we present a novel use of jerk to make the acceleration method robust to quick large motions. Jerk has been used to assess smoothness of time series data in the neuroscience and mechanical engineering fields. On the basis of our observation that subtle changes are smoother than quick large motions at temporal scale, we used jerk-based smoothness to design a jerk-aware filter that passes subtle changes only under quick large motions. By applying our filter to the acceleration method, we obtain impressive magnification results better than those obtained with state-of-the-art.

****

Defense Against Adversarial Attacks Using High-Level Representation Guided Denoiser

Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, Jun Zhu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1778-1787

Neural networks are vulnerable to adversarial examples, which poses a threat to their application in security sensitive systems. We propose high-level representation guided denoiser (HGD) as a defense for image classification. Standard denoiser suffers from the error amplification effect, in which small residual adversarial noise is progressively amplified and leads to wrong classifications. HGD o vercomes this problem by using a loss function defined as the difference between the target model's outputs activated by the clean image and denoised image. Compared with ensemble adversarial training which is the state-of-the-art defending method on large images, HGD has three advantages. First, with HGD as a defense, the target model is more robust to either white-box or black-box adversarial attacks. Second, HGD can be trained on a small subset of the images and generalizes well to other images and unseen classes. Third, HGD can be transferred to defend models other than the one guiding it. In NIPS competition on defense against adversarial attacks, our HGD solution won the first place and outperformed other models by a large margin. footnote{Code: url{https://github.com/lfz/Guided-Denoise}.}

****

Stacked Conditional Generative Adversarial Networks for Jointly Learning Shadow Detection and Shadow Removal

Jifeng Wang, Xiang Li, Jian Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1788-1797

Understanding shadows from a single image consists of two types of task in previous studies, containing shadow detection and shadow removal. In this paper, we present a multi-task perspective, which is not embraced by any existing work, to jointly learn both detection and removal in an end-to-end fashion that aims at enjoying the mutually improved benefits from each other. Our framework is based on a novel STacked Conditional Generative Adversarial Network (ST-CGAN), which is composed of two stacked CGANs, each with a generator and a discriminator. Specifically, a shadow image is fed into the first generator which produces a shadow detection mask. That shadow image, concatenated with its predicted mask, goes through the second generator in order to recover its shadow-free image consequently. In addition, the two corresponding discriminators are very likely to model higher level relationships and global scene characteristics for the detected shadow region and reconstruction via removing shadows, respectively. More importantly, for multi-task learning, our design of stacked paradigm provides a novel view which is notably different from the commonly used one as the multi-branch version. To fully evaluate the performance of our proposed framework, we construct the first large-scale benchmark with 1870 image triplets (shadow image, shadow mask image, and shadow-free image) under 135 scenes. Extensive experimental results consistently show the advantages of ST-CGAN over several representative state-of-the-art methods on two large-scale publicly available datasets and our newly released one.

****

Image Correction via Deep Reciprocating HDR Transformation

Xin Yang, Ke Xu, Yibing Song, Qiang Zhang, Xiaopeng Wei, Rynson W.H. Lau; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1798-1807

Image correction aims to adjust an input image into a visually pleasing one with the detail in the under/over exposed regions recovered. However, existing image correction methods are mainly based on image pixel operations, and attempting to recover the lost detail from these under/over exposed regions is challenging. We, therefore, revisit the image formation procedure and notice that detail is contained in the high dynamic range (HDR) light intensities(which can be perceived by human eyes) but is lost during the nonlinear imaging process by of the camera in the low dynamic range (LDR) domain. Inspired by this observation, we formulate the image correction problem as the Deep Reciprocating HDR Transformation (DRHT) process and propose a novel approach to first reconstruct the lost detail in the HDR domain and then transfer them back to the LDR image as the output image with the recovered detail preserved. To this end, we propose an end-to-end DRHT model, which contains two CNNs, one for HDR detail reconstruction and the other for LDR detail correction. Experiments on the standard benchmarks demonstrate the effectiveness of the proposed method, compared with state-of-the-art image correction methods.

****

PieAPP: Perceptual Image-Error Assessment Through Pairwise Preference

Ekta Prashnani, Hong Cai, Yasamin Mostofi, Pradeep Sen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1808-1817

The ability to estimate the perceptual error between images is an important problem in computer vision with many applications. Although it has been studied extensively, however, no method currently exists that can robustly predict visual differences like humans. Some previous approaches used hand-coded models, but they fail to model the complexity of the human visual system. Others used machine learning to train models on human-labeled datasets, but creating large, high-quality datasets is difficult because people are unable to assign consistent error labels to distorted images. In this paper, we present a new learning-based method that is the first to predict perceptual image error like human observers. Since it is much easier for people to compare two given images and identify the one more similar to a reference than to assign quality scores to each, we propose a new, large-scale dataset labeled with the probability that humans will prefer one image over another. We then train a deep-learning model using a novel, pairwise-learning framework to predict the preference of one distorted image over the other. Our key observation is that our trained network can then be used separately with only one distorted image and a reference to predict its perceptual error, without ever being trained on explicit human perceptual-error labels. The perceptual error estimated by our new metric, PieAPP, is well-correlated with human opinion. Furthermore, it significantly outperforms existing algorithms, beating the state-of-the-art by almost 3x on our test set in terms of binary error rate, while also generalizing to new kinds of distortions, unlike previous learning-based methods.

****

Normalized Cut Loss for Weakly-Supervised CNN Segmentation

Meng Tang, Abdelaziz Djelouah, Federico Perazzi, Yuri Boykov, Christopher Schroers; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1818-1827

Most recent semantic segmentation methods train deep convolutional neural networks with fully annotated masks requiring pixel-accuracy for good quality training. Common weakly-supervised approaches generate full masks from partial input (e.g. scribbles or seeds) using standard interactive segmentation methods as preprocessing. But, errors in such masks result in poorer training since standard loss functions (e.g. cross-entropy) do not distinguish seeds from potentially mislabeled other pixels. Inspired by the general ideas in semi-supervised learning, we address these problems via a new principled loss function evaluating network output with criteria standard in ``shallow'' segmentation, e.g. normalized cut. Unlike prior work, the cross entropy part of our loss evaluates only seeds where labels are known while normalized cut softly evaluates consistency of all pixels. We focus on normalized cut loss where dense Gaussian kernel is efficiently implemented in linear time by fast Bilateral filtering. Our normalized cut loss approach to segmentation brings the quality of weakly-supervised training significantly closer to fully supervised methods.

****

ISTA-Net: Interpretable Optimization-Inspired Deep Network for Image Compressive Sensing

Jian Zhang, Bernard Ghanem; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1828-1837

With the aim of developing a fast yet accurate algorithm for compressive sensing (CS) reconstruction of natural images, we combine in this paper the merits of two existing categories of CS methods: the structure insights of traditional optimization-based methods and the performance/speed of recent network-based ones. Specifically, we propose a novel structured deep network, dubbed ISTA-Net, which is inspired by the Iterative Shrinkage-Thresholding Algorithm (ISTA) for optimizing a general L1 norm CS reconstruction model. To cast ISTA into deep network form, we develop an effective strategy to solve the proximal mapping associated with the sparsity-inducing regularizer using nonlinear transforms. All the parameters in ISTA-Net (e.g. nonlinear transforms, shrinkage thresholds, step sizes, e tc.) are learned end-to-end, rather than being hand-crafted. Moreover, considering that the residuals of natural images are more compressible, an enhanced version of ISTA-Net in the residual domain, dubbed ISTA-Net+, is derived to further improve CS reconstruction. Extensive CS experiments demonstrate that the proposed ISTA-Nets outperform existing state-of-the-art optimization-based and network-based CS methods by large margins, while maintaining fast computational speed.

****

Fast End-to-End Trainable Guided Filter

Huikai Wu, Shuai Zheng, Junge Zhang, Kaiqi Huang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1838-1847

Image processing and pixel-wise dense prediction have been advanced by harnessing the capabilities of deep learning. One central issue of deep learning is the limited capacity to handle joint upsampling. We present a deep learning building block for joint upsampling, namely guided filtering layer. This layer aims at efficiently generating the high-resolution output given the corresponding low-resolution one and a high-resolution guidance map. The proposed layer is composed of a guided filter, which is reformulated as a fully differentiable block. To this end, we show that a guided filter can be expressed as a group of spatial varying linear transformation matrices. This layer could be integrated with the convolutional neural networks (CNNs) and jointly optimized through end-to-end training. To further take advantage of end-to-end training, we plug in a trainable transformation function that generates task-specific guidance maps. By integrating the CNNs and the proposed layer, we form deep guided filtering networks. The proposed networks are evaluated on five advanced image processing tasks. Experiments on MIT-Adobe FiveK Dataset demonstrate that the proposed approach runs 10-100 times faster and achieves the state-of-the-art performance. We also show that the proposed guided filtering layer helps to improve the performance of multiple pixel-wise dense prediction tasks.

****

Disentangling Structure and Aesthetics for Style-Aware Image Completion

Andrew Gilbert, John Collomosse, Hailin Jin, Brian Price; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1848-1856

Content-aware image completion or in-painting is a fundamental tool for the correction of defects or removal of objects in images. We propose a non-parametric in-painting algorithm that enforces both structural and aesthetic (style) consistency within the resulting image. Our contributions are two-fold: 1) we explicitly disentangle image structure and style during patch search and selection to ensure a visually consistent look and feel within the target image; 2) we perform adaptive stylization of patches to conform the aesthetics of selected patches to the target image, so harmonising the integration of selected patches into the final composition. We show that explicit consideration of visual style during in-painting delivers excellent qualitative and quantitative results across the varied image styles and content, over the Places2 photographic dataset and a challenging new in-painting dataset of artwork derived from BAM!

****

Learning a Discriminative Feature Network for Semantic Segmentation

Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, Nong Sang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1857-1866

Most existing methods of semantic segmentation still suffer from two aspects of challenges: intra-class inconsistency and inter-class indistinction. To tackle these two problems, we propose a Discriminative Feature Network (DFN), which contains two sub-networks: Smooth Network and Border Network. Specifically, to handle the intra-class inconsistency problem, we specially design a Smooth Network with Channel Attention Block and global average pooling to select the more discriminative features. Furthermore, we propose a Border Network to make the bilateral features of boundary distinguishable with deep semantic boundary supervision. Based on our proposed DFN, we achieve state-of-the-art performance 86.2% mean IOU on PASCAL VOC 2012 and 80.3% mean IOU on Cityscapes dataset.

****

Kernelized Subspace Pooling for Deep Local Descriptors

Xing Wei, Yue Zhang, Yihong Gong, Nanning Zheng; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1867-1875

Representing local image patches in an invariant and discriminative manner is an active research topic in computer vision. It has recently been demonstrated that local feature learning based on deep Convolutional Neural Network (CNN) can significantly improve the matching performance. Previous works on learning such descriptors have focused on developing various loss functions, regularizations and data mining strategies to learn discriminative CNN representations. Such methods, however, have little analysis on how to increase geometric invariance of their generated descriptors. In this paper, we propose a descriptor that has both highly invariant and discriminative power. The abilities come from a novel pooling method, dubbed Subspace Pooling (SP) which is invariant to a range of geometric deformations. To further increase the discriminative power of our descriptor, we propose a simple distance kernel integrated to the marginal triplet loss that helps to focus on hard examples in CNN training. Finally, we show that by combining SP with the projection distance metric, the generated feature descriptor is equivalent to that of the Bilinear CNN model, but outperforms the latter with much lower memory and computation consumptions. The proposed method is simple, easy to understand and achieves good performance. Experimental results on several patch matching benchmarks show that our method outperforms the state-of-the-arts significantly.

****

pOSE: Pseudo Object Space Error for Initialization-Free Bundle Adjustment

Je Hyeong Hong, Christopher Zach; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1876-1885

Bundle adjustment is a nonlinear refinement method for camera poses and 3D structure requiring sufficiently good initialization. In recent years, it was experimentally observed that useful minima can be reached even from arbitrary initialization for affine bundle adjustment problems (and fixed-rank matrix factorization instances in general). The key success factor lies in the use of the variable projection (VarPro) method, which is known to have a wide basin of convergence for such problems. In this paper, we propose the Pseudo Object Space Error (pOSE), which is an objective with cameras represented as a hybrid between the affine and projective models. This formulation allows us to obtain 3D reconstructions that are close to the true projective reconstructions while retaining a bilinear problem structure suitable for the VarPro method. Experimental results show that using pOSE has a high success rate to yield faithful 3D reconstructions from random initializations, taking one step towards initialization-free structure from motion.

****

Deformable Shape Completion With Graph Convolutional Autoencoders

Or Litany, Alex Bronstein, Michael Bronstein, Ameesh Makadia; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1886-1895

The availability of affordable and portable depth sensors has made scanning objects and people simpler than ever. However, dealing with occlusions and missing parts is still a significant challenge. The problem of reconstructing a (possibly non-rigidly moving) 3D object from a single or multiple partial scans has received increasing attention in recent years. In this work, we propose a novel learningbased method for the completion of partial shapes. Unlike the majority of existing approaches, our method focuses on objects that can undergo non-rigid deformations. The core of our method is a variational autoencoder with graph convolutional operations that learns a latent space for complete realistic shapes. At inference, we optimize to find the representation in this latent space that best fits the generated shape to the known partial input. The completed shape exhibits a realistic appearance on the unknown part. We show promising results towards the completion of synthetic and real scans of human body and face meshes exhibiting different styles of articulation and partiality.

****

Learning From Millions of 3D Scans for Large-Scale 3D Face Recognition

Syed Zulqarnain Gilani, Ajmal Mian; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1896-1905

Deep networks trained on millions of facial images are believed to be closely approaching human-level performance in face recognition. However, open world face recognition still remains a challenge. Although, 3D face recognition has an inherent edge over its 2D counterpart, it has not benefited from the recent developments in deep learning due to the unavailability of large training as well as large test datasets. Recognition accuracies have already saturated on existing 3D face datasets due to their small gallery sizes. Unlike 2D photographs, 3D facial scans cannot be sourced from the web causing a bottleneck in the development of deep 3D face recognition networks and datasets. In this backdrop, we propose a method for generating a large corpus of labeled 3D face identities and their multiple instances for training and a protocol for merging the most challenging existing 3D datasets for testing. We also propose the first deep CNN model designed specifically for 3D face recognition and trained on 3.1 Million 3D facial scans of 100K identities. Our test dataset comprises 1,853 identities with a single 3D scan in the gallery and another 31K scans as probes, which is several orders of magnitude larger than existing ones. Without fine tuning on this dataset, our network already outperforms state of the art face recognition by over 10%. We fine tune our network on the gallery set to perform end-to-end large scale 3D face recognition which further improves accuracy. Finally, we show the efficacy of our method for the open world face recognition problem.

****

CarFusion: Combining Point Tracking and Part Detection for Dynamic 3D Reconstruction of Vehicles

N. Dinesh Reddy, Minh Vo, Srinivasa G. Narasimhan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1906-1915

Despite significant research in the area, reconstruction of multiple dynamic rigid objects (eg. vehicles) observed from   wide-baseline, uncalibrated and unsynchronized cameras, remains hard. On one hand, feature tracking works well within each view but is hard to correspond across multiple cameras with limited overlap in fields of view or due to occlusions. On the other hand, advances in deep learning have resulted in strong detectors that work across different viewpoints but are still not precise enough for triangulation-based reconstruction. In this work, we develop a framework to fuse both the single-view feature tracks and multi-view detected part locations to significantly improve the detection, localization and reconstruction of moving vehicles, even in the presence of strong occlusions. We demonstrate our framework at a busy traffic intersection by reconstructing over 62 vehicles passing within a 3-minute window. We evaluate the different components within our framework and compare to alternate approaches such as reconstruction using tracking-by-detection.

****

Deep Material-Aware Cross-Spectral Stereo Matching

Tiancheng Zhi, Bernardo R. Pires, Martial Hebert, Srinivasa G. Narasimhan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1916-1925

Cross-spectral imaging provides strong benefits for recognition and detection tasks. Often, multiple cameras are used for cross-spectral imaging, thus requiring image alignment, or disparity estimation in a stereo setting. Increasingly, multi-camera cross-spectral systems are embedded in active RGBD devices (e.g. RGB-NIR cameras in Kinect and iPhone X). Hence, stereo matching also provides an opportunity to obtain depth without an active projector source. However, matching images from different spectral bands is challenging because of large appearance variations. We develop a novel deep learning framework to simultaneously transform images across spectral bands and estimate disparity. A material-aware loss function is incorporated within the disparity prediction network to handle regions with unreliable matching such as light sources, glass windshields and glossy surfaces. No depth supervision is required by our method. To evaluate our method, we used a vehicle-mounted RGB-NIR stereo system to collect 13.7 hours of video data across a range of areas in and around a city. Experiments show that our method achieves strong performance and reaches real-time speed.

****

Augmenting Crowd-Sourced 3D Reconstructions Using Semantic Detections

True Price, Johannes L. Schönberger, Zhen Wei, Marc Pollefeys, Jan-Michael Frahm; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1926-1935

Image-based 3D reconstruction for Internet photo collections has become a robust technology to produce impressive virtual representations of real-world scenes. However, several fundamental challenges remain for Structure-from-Motion (SfM) pipelines, namely: the placement and reconstruction of transient objects only observed in single views, estimating the absolute scale of the scene, and (suprisingly often) recovering ground surfaces in the scene. We propose a method to jointly address these remaining open problems of SfM. In particular, we focus on detecting people in individual images and accurately placing them into an existing 3D model. As part of this placement, our method also estimates the absolute scale of the scene from object semantics, which in this case constitutes the height distribution of the population. Further, we obtain a smooth approximation of the ground surface and recover the gravity vector of the scene directly from the individual person detections. We demonstrate the results of our approach on a number of unordered Internet photo collections, and we quantitatively evaluate the obtained absolute scene scales.

****

Matryoshka Networks: Predicting 3D Geometry via Nested Shape Layers

Stephan R. Richter, Stefan Roth; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1936-1944

In this paper, we develop novel, efficient 2D encodings for 3D geometry, which enable reconstructing full 3D shapes from a single image at high resolution. The key idea is to pose 3D shape reconstruction as a 2D prediction problem. To that end, we first develop a simple baseline network that predicts entire voxel tubes at each pixel of a reference view. By leveraging well-proven architectures for 2D pixel-prediction tasks, we attain state-of-the-art results, clearly outperforming purely voxel-based approaches. We scale this baseline to higher resolutions by proposing a memory-efficient shape encoding, which recursively decomposes a 3D shape into nested shape layers, similar to the pieces of a Matryoshka doll. This allows reconstructing highly detailed shapes with complex topology, as demonstrated in extensive experiments; we clearly outperform previous octree-based approaches despite having a much simpler architecture using standard network components. Our Matryoshka networks further enable reconstructing shapes from IDs or shape similarity, as well as shape sampling.

****

Triplet-Center Loss for Multi-View 3D Object Retrieval

Xinwei He, Yang Zhou, Zhichao Zhou, Song Bai, Xiang Bai; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1945-1954

Most existing 3D object recognition algorithms focus on leveraging the strong discriminative power of deep learning models with softmax loss for the classification of 3D data, while learning discriminative features with deep metric learning for 3D object retrieval is more or less neglected. In the paper, we study variants of deep metric learning losses for 3D object retrieval, which did not receive enough attention from this area. First , two kinds of representative losses, triplet loss and center loss, are introduced which could learn more discriminative features than traditional classification loss. Then we propose a novel loss named triplet-center loss, which can further enhance the discriminative power of the features. The proposed triplet-center loss learns a center for each class and requires that the distances between samples and centers from the same class are closer than those from different classes. Extensive experimental results on two popular 3D object retrieval benchmarks and two widely-adopted sketch-based 3D shape retrieval benchmarks consistently demonstrate the effectiveness of our proposed loss, and significant improvements have been achieved compared to the state-of-the-arts.

****

Learning 3D Shape Completion From Laser Scan Data With Weak Supervision

David Stutz, Andreas Geiger; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1955-1964

3D shape completion from partial point clouds is a fundamental problem in computer vision and computer graphics. Recent approaches can be characterized as either data-driven or learning-based. Data-driven approaches rely on a shape model whose parameters are optimized to fit the observations. Learning-based approaches, in contrast, avoid the expensive optimization step and instead directly predict the complete shape from the incomplete observations using deep neural networks. However, full supervision is required which is often not available in practice. In this work, we propose a weakly-supervised learning-based approach to 3D shape completion which neither requires slow optimization nor direct supervision. While we also learn a shape prior on synthetic data, we amortize, i.e., learn, maximum likelihood fitting using deep neural networks resulting in efficient shape completion without sacrificing accuracy. Tackling 3D shape completion of cars on ShapeNet and KITTI, we demonstrate that the proposed amortized maximum likelihood approach is able to compete with a fully supervised baseline and a state-of-the-art data-driven approach while being significantly faster. On ModelNet, we additionally show that the approach is able to generalize to other object categories as well.

****

End-to-End Learning of Keypoint Detector and Descriptor for Pose Invariant 3D Matching

Georgios Georgakis, Srikrishna Karanam, Ziyan Wu, Jan Ernst, Jana Košecká; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1965-1973

Finding correspondences between images or 3D scans is at the heart of many computer vision and image retrieval applications and is often enabled by matching local keypoint descriptors. Various learning approaches have been applied in the past to different stages of the matching pipeline, considering detection, description, or metric learning objectives. These objectives were typically addressed separately and most previous work has focused on image data. This paper proposes an end-to-end learning framework for keypoint detection and its representation (descriptor) for 3D depth maps or 3D scans, where the two can be jointly optimized towards task-specific objectives without a need for separate annotations. We employ a Siamese architecture augmented by a sampling layer and a novel score loss function which in turn affects the selection of region proposals. The positive and negative examples are obtained automatically by sampling corresponding region proposals based on their consistency with known 3D pose labels. Matching experiments with depth data on multiple benchmark datasets demonstrate the efficacy of the proposed approach, showing significant improvements over state-of-the-art methods.

****

ICE-BA: Incremental, Consistent and Efficient Bundle Adjustment for Visual-Inertial SLAM

Haomin Liu, Mingyu Chen, Guofeng Zhang, Hujun Bao, Yingze Bao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1974-1982

Modern visual-inertial SLAM (VI-SLAM) achieves higher accuracy and robustness than pure visual SLAM, thanks to the complementariness of visual features and inertial measurements. However, jointly using visual and inertial measurements to optimize SLAM objective functions is a problem of high computational complexity. In many VI-SLAM applications, the conventional optimization solvers can only use a very limited number of recent measurements for real time pose estimation, at the cost of suboptimal localization accuracy. In this work, we renovate the numerical solver for VI-SLAM. Compared to conventional solvers, our proposal provides an exact solution with significantly higher computational efficiency. Our solve r allows us to use remarkably larger number of measurements to achieve higher accuracy and robustness. Furthermore, our method resolves the global consistency problem that is unaddressed by many state-of-the-art SLAM systems: to guarantee the minimization of re-projection function and inertial constraint function during loop closure. Experiments demonstrate our novel formulation renders lower localization error and more than 10x speedup compared to alternatives. We release the source code of our implementation to benefit the community.

****

GeoNet: Unsupervised Learning of Dense Depth, Optical Flow and Camera Pose

Zhichao Yin, Jianping Shi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1983-1992

We propose GeoNet, a jointly unsupervised learning framework for monocular depth, optical flow and ego-motion estimation from videos. The three components are coupled by the nature of 3D scene geometry, jointly learned by our framework in an end-to-end manner. Specifically, geometric relationships are extracted over the predictions of individual modules and then combined as an image reconstruction loss, reasoning about static and dynamic scene parts separately. Furthermore, we propose an adaptive geometric consistency loss to increase robustness towards outliers and non-Lambertian regions, which resolves occlusions and texture ambiguities effectively. Experimentation on the KITTI driving dataset reveals that our scheme achieves state-of-the-art results in all of the three tasks, performing better than previously unsupervised methods and comparably with supervised ones.

****

Radially-Distorted Conjugate Translations

James Pritts, Zuzana Kukelova, Viktor Larsson, Ond■ej Chum; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1993-2001

This paper introduces the first minimal solvers that jointly solve for affine-rectification and radial lens distortion from coplanar repeated patterns. Even with imagery from moderately distorted lenses, plane rectification using the pinhole camera model is inaccurate or invalid. The proposed solvers incorporate lens distortion into the camera model and extend accurate rectification to wide-angle imagery, which is now common from consumer cameras. The solvers are derived from constraints induced by the conjugate translations of an imaged scene plane, which are integrated with the division model for radial lens distortion. The hidden-variable trick with ideal saturation is used to reformulate the constraints so that the solvers generated by the Gr{\"o}bner-basis method are stable, small and fast. The proposed solvers are used in a RANSAC-based estimator. Rectification and lens distortion are recovered from either one conjugately translated affine-covariant feature or two independently translated similarity-covariant features. Experiments confirm that RANSAC accurately estimates the rectification and radial distortion with very few iterations. The proposed solvers are evaluated against the state-of-the-art for affine rectification and radial distortion estimation.

****

Deep Ordinal Regression Network for Monocular Depth Estimation

Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, Dacheng Tao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2002-2011

Monocular depth estimation, which plays a crucial role in understanding 3D scene geometry, is an ill-posed prob- lem. Recent methods have gained significant improvement by exploring image-level information and hierarchical features from deep convolutional neural networks (DCNNs). These methods model depth estimation as a regression problem and train the regression networks by minimizing mean squared error, which suffers from slow convergence and unsatisfactory local solutions. Besides, existing depth estimation networks employ repeated spatial pooling operations, resulting in undesirable low-resolution feature maps. To obtain high-resolution depth maps, skip-connections or multi- layer deconvolution networks are required, which complicates network training and consumes much more computatio

ns. To eliminate or at least largely reduce these problems, we introduce a spacing-increasing discretization (SID) strategy to discretize depth and recast depth network learning as an ordinal regression problem. By training the network using an ordinary regression loss, our method achieves much higher accuracy and faster convergence in synch. Furthermore, we adopt a multi-scale network structure which avoids unnecessary spatial pooling and captures multi-scale information in parallel. The proposed deep ordinal regression network (DORN) achieves state-of-the-art results on three challenging benchmarks, i.e., KITTI [16], Make3D [49], and NYU Depth v2 [41], and outperforms existing methods by a large margin.
****

Analytical Modeling of Vanishing Points and Curves in Catadioptric Cameras

Pedro Miraldo, Francisco Eiras, Srikumar Ramalingam; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2012-2021

Vanishing points and vanishing lines are classical geometrical concepts in perspective cameras that have a lineage dating back to 3 centuries. A vanishing point is a point on the image space where parallel lines in 3D space appear to converge, whereas a vanishing line passes through 2 or more vanishing points. While such concepts are simple and intuitive in perspective cameras, their counterparts in catadioptric cameras (obtained using mirrors and lenses) are more involved. For example, lines in the 3D space map to higher degree curves in catadioptric cameras. The projection of a set of 3D parallel lines converges on a single point in perspective images, whereas they converge to more than one point in catadioptric cameras. To the best of our knowledge, we are not aware of any systematic development of analytical models for vanishing points and vanishing curves in different types of catadioptric cameras. In this paper, we derive parametric equations for vanishing points and vanishing curves using the calibration parameters, mirror shape coefficients, and direction vectors of parallel lines in 3D space. We show compelling experimental results on vanishing point estimation and absolute pose estimation for a wide variety of catadioptric cameras in both simulations and real experiments.
****

Learning Depth From Monocular Videos Using Direct Methods

Chaoyang Wang, José Miguel Buenaposada, Rui Zhu, Simon Lucey; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2022-2030

The ability to predict depth from a single image - using recent advances in CNNs - is of increasing interest to the vision community. Unsupervised strategies to learning are particularly appealing as they can utilize much larger and varied monocular video datasets during learning without the need for ground truth depth or stereo. In previous works, separate pose and depth CNN predictors had to be determined such that their joint outputs minimized the photometric error. Inspired by recent advances in direct visual odometry (DVO), we argue that the depth CNN predictor can be learned without a pose CNN predictor. Further, we demonstrate empirically that incorporation of a differentiable implementation of DVO, along with a novel depth normalization strategy - substantially improves performance over state of the art that use monocular videos for training.
****

Salience Guided Depth Calibration for Perceptually Optimized Compressive Light Field 3D Display

Shizheng Wang, Wenjuan Liao, Phil Surman, Zhigang Tu, Yuanjin Zheng, Junsong Yuan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2031-2040

Multi-layer light field displays are a type of computational three-dimensional (3D) display which has recently gained increasing interest for its holographic-like effect and natural compatibility with 2D displays. However, the major shortcoming, depth limitation, still cannot be overcome in the traditional light field modeling and reconstruction based on multi-layer liquid crystal displays (LCDs). Considering this disadvantage, our paper incorporates a salience guided depth optimization over a limited display range to calibrate the displayed depth and present the maximum area of salience region for multi-layer light field display. D ifferent from previously reported cascaded light field displays that use the fixed initialization plane as the depth center of display content, our method automatically calibrates the depth initialization based on the salience results derived from the proposed contrast enhanced salience detection method. Experiments demonstrate that the proposed method provides a promising advantage in visual perception for the compressive light field displays from both software simulation and prototype demonstration.

****

MegaDepth: Learning Single-View Depth Prediction From Internet Photos
Zhengqi Li, Noah Snavely; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2041-2050

Single-view depth prediction is a fundamental problem in computer vision. Recently, deep learning methods have led to significant progress, but such methods are limited by the available training data. Current datasets based on 3D sensors have key limitations, including indoor-only images (NYU), small numbers of training examples (Make3D), and sparse sampling (KITTI). We propose to use multi-view Internet photo collections, a virtually unlimited data source, to generate training data via modern structure-from-motion and multi-view stereo (MVS) methods, and present a large depth dataset called MegaDepth based on this idea. Data derived from MVS comes with its own challenges, including noise and unreconstructable objects. We address these challenges with new data cleaning methods, as well as automatically augmenting our data with ordinal depth relations generated using semantic segmentation. We validate the use of large amounts of Internet data by showing that models trained on MegaDepth exhibit strong generalization—not only to novel scenes, but also to other diverse datasets including Make3D, KITTI, and DIW, even when no images from those datasets are seen during training.

****

LayoutNet: Reconstructing the 3D Room Layout From a Single RGB Image
Chuhang Zou, Alex Colburn, Qi Shan, Derek Hoiem; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2051-2059

We propose an algorithm to predict room layout from a single image that generalizes across panoramas and perspective images, cuboid layouts and more general layouts (e.g. "L"-shape room). Our method operates directly on the panoramic image, rather than decomposing into perspective images as do recent works. Our network architecture is similar to that of RoomNet, but we show improvements due to aligning the image based on vanishing points, predicting multiple layout elements (corners, boundaries, size and translation), and fitting a constrained Manhattan layout to the resulting predictions. Our method compares well in speed and accuracy to other existing work on panoramas, achieves among the best accuracy for perspective images, and can handle both cuboid-shaped and more general Manhattan layouts.

****

CBMV: A Coalesced Bidirectional Matching Volume for Disparity Estimation
Konstantinos Batsos, Changjiang Cai, Philippos Mordohai; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2060-2069

Recently, there has been a paradigm shift in stereo matching with learning-based methods achieving the best results on all popular benchmarks. The success of these methods is due to the availability of training data with ground truth; training learning-based systems on these datasets has allowed them to surpass the accuracy of conventional approaches based on heuristics and assumptions. Many of these assumptions, however, had been validated extensively and hold for the majority of possible inputs. In this paper, we generate a matching volume leveraging both data with ground truth and conventional wisdom. We accomplish this by coalescing diverse evidence from a bidirectional matching process via random forest classifiers. We show that the resulting matching volume estimation method achieves similar accuracy to purely data-driven alternatives on benchmarks and that it generalizes to unseen data much better. In fact, the results we submitted to the KITTI benchmarks were generated using a classifier trained on the Middlebury dataset.

****

Zoom and Learn: Generalizing Deep Stereo Matching to Novel Domains

Jiahao Pang, Wenxiu Sun, Chengxi Yang, Jimmy Ren, Ruichao Xiao, Jin Zeng, Liang Lin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2070-2079

Despite the recent success of stereo matching with convolutional neural networks (CNNs), it remains arduous to generalize a pre-trained deep stereo model to a novel domain. A major difficulty is to collect accurate ground-truth disparities for stereo pairs in the target domain. In this work, we propose a self-adaptation approach for CNN training, utilizing both synthetic training data (with ground-truth disparities) and stereo pairs in the new domain (without ground-truths). Our method is driven by two empirical observations. By feeding real stereo pairs of different domains to stereo models pre-trained with synthetic data, we see that: i) a pre-trained model does not generalize well to the new domain, producing artifacts at boundaries and ill-posed regions; however, ii) feeding an up-sampled stereo pair leads to a disparity map with extra details. To avoid i) while exploiting ii), we formulate an iterative optimization problem with graph Laplacian regularization. At each iteration, the CNN adapts itself better to the new domain: we let the CNN learn its own higher-resolution output; at the meanwhile, a graph Laplacian regularization is imposed to discriminatively keep the desired edges while smoothing out the artifacts. We demonstrate the effectiveness of our method in two domains: daily scenes collected by smartphone cameras, and street views captured in a driving car.

****

Exploring Disentangled Feature Representation Beyond Face Identification

Yu Liu, Fangyin Wei, Jing Shao, Lu Sheng, Junjie Yan, Xiaogang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2080-2089

This paper proposes learning disentangled but complementary face features with a minimal supervision by face identification. Specifically, we construct an identity Distilling and Dispelling Auto-Encoder (D^2AE) framework that adversarially learns the identity-distilled features for identity verification and the identity-dispelled features to fool the verification system. Thanks to the design of two-stream cues, the learned disentangled features represent not only the identity or attribute but the complete input image. Comprehensive evaluations further demonstrate that the proposed features not only preserve state-of-the-art identity verification performance on LFW, but also acquire comparable discriminative power for face attribute recognition on CelebA and LFWA. Moreover, the proposed system is ready to semantically control the face generation/editing based on various identities and attributes in an unsupervised manner.

****

Learning Facial Action Units From Web Images With Scalable Weakly Supervised Clustering

Kaili Zhao, Wen-Sheng Chu, Aleix M. Martinez; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2090-2099

We present a scalable weakly supervised clustering approach to learn facial action units (AUs) from large, freely available web images. Unlike most existing methods (e.g., CNNs) that rely on fully annotated data, our method exploits web images with inaccurate annotations. Specifically, we derive a weakly-supervised spectral algorithm that learns an embedding space to couple image appearance and semantics. The algorithm has efficient gradient update, and scales up to large quantities of images with a stochastic extension. With the learned embedding space, we adopt rank-order clustering to identify groups of visually and semantically similar images, and re-annotate these groups for training AU classifiers. Evaluation on the 1 millon EmotioNet dataset demonstrates the effectiveness of our approach: (1) our learned annotations reach on average 91.3% agreement with human annotations on 7 common AUs, (2) classifiers trained with re-annotated images perform comparably to, sometimes even better than, its supervised CNN-based counterpart, and (3) our method offers intuitive outlier/noise pruning instead of forcing one annotation to every image. Code is available.

****

Human Pose Estimation With Parsing Induced Learner

Xuecheng Nie, Jiashi Feng, Yiming Zuo, Shuicheng Yan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2100-2108

Human pose estimation still faces various difficulties in challenging scenarios. Human parsing, as a closely related task, can provide valuable cues for better pose estimation, which however has not been fully exploited. In this paper, we propose a novel Parsing Induced Learner to exploit parsing information to effectively assist pose estimation by learning to fast adapt the base pose estimation model. The proposed Parsing Induced Learner is composed of a parsing encoder and a pose model parameter adapter, which together learn to predict dynamic parameters of the pose model to extract complementary useful features for more accurate pose estimation. Comprehensive experiments on benchmarks LIP and extended PASCAL-Person-Part show that the proposed Parsing Induced Learner can improve performance of both single- and multi-person pose estimation to new state-of-the-art. Cross-dataset experiments also show that the proposed Parsing Induced Learner from LIP dataset can accelerate learning of a human pose estimation model on MPII benchmark in addition to achieving outperforming performance.

****

Multi-Level Factorisation Net for Person Re-Identification

Xiaobin Chang, Timothy M. Hospedales, Tao Xiang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2109-2118

Key to effective person re-identification (Re-ID) is modelling discriminative and view-invariant factors of person appearance at both high and low semantic levels. Recently developed deep Re-ID models either learn a holistic single semantic level feature representation and/or require laborious human annotation of these factors as attributes. We propose Multi-Level Factorisation Net (MLFN), a novel network architecture that factorises the visual appearance of a person into latent discriminative factors at multiple semantic levels without manual annotation. MLFN is composed of multiple stacked blocks. Each block contains multiple factor modules to model latent factors at a specific level, and factor selection modules that dynamically select the factor modules to interpret the content of each input image. The outputs of the factor selection modules also provide a compact latent factor descriptor that is complementary to the conventional deeply learned features. MLFN achieves state-of-the-art results on three Re-ID datasets, as well as compelling results on the general object categorisation CIFAR-100 dataset.

****

Attention-Aware Compositional Network for Person Re-Identification

Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, Wanli Ouyang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2119-2128

Person re-identification (ReID) is to identify pedestrians observed from different camera views based on visual appearance. It is a challenging task due to large pose variations, complex background clutters and severe occlusions. Recently, human pose estimation by predicting joint locations was largely improved in accuracy. It is reasonable to use pose estimation results for handling pose variations and background clutters, and such attempts have obtained great improvement in ReID performance. However, we argue that the pose information was not well utilized and hasn’t yet been fully exploited for person ReID. In this work, we introduce a novel framework called Attention-Aware Compositional Network (AACN) for person ReID. AACN consists of two main components: Pose-guided Part Attention (PPA) and Attention-aware Feature Composition (AFC). PPA is learned and applied to mask out undesirable background features in pedestrian feature maps. Furthermore, pose-guided visibility scores are estimated for body parts to deal with part occlusion in the proposed AFC module. Extensive experiments with ablation analysis show the effectiveness of our method, and state-of-the-art results are achieved on several public datasets, including Market-1501, CUHK03, CUHK01, SenseReID, CUHK03-NP and DukeMTMC-reID.

****

Look at Boundary: A Boundary-Aware Face Alignment Algorithm
Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, Qiang Zhou; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2129-2138

We present a novel boundary-aware face alignment algorithm by utilising boundary lines as the geometric structure of a human face to help facial landmark localisation. Unlike the conventional heatmap based method and regression based method, our approach derives face landmarks from boundary lines which remove the ambiguities in the landmark definition. Three questions are explored and answered by this work: 1. Why using boundary? 2. How to use boundary? 3. What is the relationship between boundary estimation and landmarks localisation? Our boundary-aware face alignment algorithm achieves 3.49% mean error on 300-W Fullset, which outperforms state-of-the-art methods by a large margin. Our method can also easily integrate information from other datasets. By utilising boundary information of 300-W dataset, our method achieves 3.92% mean error with 0.39% failure rate on COFW dataset, and 1.25% mean error on AFLW-Full dataset. Moreover, we propose a new dataset WFLW to unify training and testing across different factors, including poses, expressions, illuminations, makeups, occlusions, and blurriness. Dataset and model are publicly available at https://wywu.github.io/projects/LAB/LAB.html

****

Demo2Vec: Reasoning Object Affordances From Online Videos
Kuan Fang, Te-Lin Wu, Daniel Yang, Silvio Savarese, Joseph J. Lim; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2139-2147

Watching expert demonstrations is an important way for humans and robots to reason about affordances of unseen objects. In this paper, we consider the problem of reasoning object affordances through the feature embedding of demonstration videos. We design the Demo2Vec model which learns to extract embedded vectors of demonstration videos and predicts the interaction region and the action label on a target image of the same object. We introduce the Online Product Review dataset for Affordance (OPRA) by collecting and labeling diverse YouTube product review videos. Our Demo2Vec model outperforms various recurrent neural network baselines on the collected dataset.

****

Monocular 3D Pose and Shape Estimation of Multiple People in Natural Scenes - The Importance of Multiple Scene Constraints
Andrei Zanfir, Elisabeta Marinoiu, Cristian Sminchisescu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2148-2157

Human sensing has greatly benefited from recent advances in deep learning, parametric human modeling, and large scale 2d and 3d datasets. However, existing 3d models make strong assumptions about the scene, considering either a single person per image, full views of the person, a simple background or many cameras. In this paper, we leverage state-of-the-art deep multi-task neural networks and parametric human and scene modeling, towards a fully automatic monocular visual sensing system for multiple interacting people, which (i) infers the 2d and 3d pose and shape of multiple people from a single image, relying on detailed semantic representations at both model and image level, to guide a combined optimization with feedforward and feedback components, (ii) automatically integrates scene constraints including ground plane support and simultaneous volume occupancy by multiple people, and (iii) extends the single image model to video by optimally solving the temporal person assignment problem and imposing coherent temporal pose and motion reconstructions while preserving image alignment fidelity. We perform experiments on both single and multi-person datasets, and systematically evaluate each component of the model, showing improved performance and extensive multiple human sensing capability. We also apply our method to images with multiple people, severe occlusions and diverse backgrounds captured in challenging natural scenes, and obtain results of good perceptual quality.

****

3D Human Sensing, Action and Emotion Recognition in Robot Assisted Therapy of Children With Autism

Elisabeta Marinoiu, Mihai Zanfir, Vlad Olaru, Cristian Sminchisescu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2158-2167

We introduce new, fine-grained action and emotion recognition tasks defined on non-staged videos, recorded during robot-assisted therapy sessions of children with autism. The tasks present several challenges: a large dataset with long videos, a large number of highly variable actions, children that are only partially visible, have different ages and may show unpredictable behaviour, as well as non-standard camera viewpoints. We investigate how state-of-the-art 3d human pose reconstruction methods perform on the newly introduced tasks and propose extensions to adapt them to deal with these challenges. We also analyze multiple approaches in action and emotion recognition from 3d human pose data, establish several baselines, and discuss results and their implications in the context of child-robot interaction.

****

Facial Expression Recognition by De-Expression Residue Learning

Huiyuan Yang, Umur Ciftci, Lijun Yin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2168-2177

A facial expression is a combination of an expressive component and a neutral component of a person. In this paper, we propose to recognize facial expressions by extracting information of the expressive component through a de-expression learning procedure, called De-expression Residue Learning (DeRL). First, a generative model is trained by cGAN. This model generates the corresponding neutral face image for any input face image. We call this procedure de-expression because the expressive information is filtered out by the generative model; however, the expressive information is still recorded in the intermediate layers. Given the neutral face image, unlike previous works using pixel-level or feature-level difference for facial expression classification, our new method learns the deposition (or residue) that remains in the intermediate layers of the generative model. Such a residue is essential as it contains the expressive component deposited in the generative model from any input facial expression images. Seven public facial expression databases are employed in our experiments. With two databases (BU-4DFE and BP4D-spontaneous) for pre-training, the DeRL method has been evaluated on five databases, CK+, Oulu-CASIA, MMI, BU- 3DFE, and BP4D+. The experimental results demonstrate the superior performance of the proposed method.

****

A Causal And-Or Graph Model for Visibility Fluent Reasoning in Tracking Interacting Objects

Yuanlu Xu, Lei Qin, Xiaobai Liu, Jianwen Xie, Song-Chun Zhu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2178-2187

Tracking humans that are interacting with the other subjects or environment remains unsolved in visual tracking, because the visibility of the human of interests in videos is unknown and might vary over time. In particular, it is still difficult for state-of-the-art human trackers to recover complete human trajectories in crowded scenes with frequent human interactions. In this work, we consider the visibility status of a subject as a fluent variable, whose change is mostly attributed to the subject's interaction with the surrounding, e.g., crossing behind another object, entering a building, or getting into a vehicle, etc. We introduce a Causal And-Or Graph (C-AOG) to represent the causal-effect relations between an object's visibility fluent and its activities, and develop a probabilistic graph model to jointly reason the visibility fluent change (e.g., from visible to invisible) and track humans in videos. We formulate this joint task as an iterative search of a feasible causal graph structure that enables fast search algorithm, e.g., dynamic programming method. We apply the proposed method on challenging video sequences to evaluate its capabilities of estimating visibility fluent changes of subjects and tracking subjects of interests over time. Results with comparisons demonstrate that our method outperforms the alternative trackers a nd can recover complete trajectories of humans in complicated scenarios with frequent human interactions.

****

Weakly Supervised Facial Action Unit Recognition Through Adversarial Training

Guozhu Peng, Shangfei Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2188-2196

Current works on facial action unit (AU) recognition typically require fully AU-annotated facial images for supervised AU classifier training. AU annotation is a time-consuming, expensive, and error-prone process. While AUs are hard to annotate, facial expression is relatively easy to label. Furthermore, there exist strong probabilistic dependencies between expressions and AUs as well as dependencies among AUs. Such dependencies are referred to as domain knowledge. In this paper, we propose a novel AU recognition method that learns AU classifiers from domain knowledge and expression-annotated facial images through adversarial training. Specifically, we first generate pseudo AU labels according to the probabilistic dependencies between expressions and AUs as well as correlations among AUs summarized from domain knowledge. Then we propose a weakly supervised AU recognition method via an adversarial process, in which we simultaneously train two models: a recognition model R, which learns AU classifiers, and a discrimination model D, which estimates the probability that AU labels generated from domain knowledge rather than the recognized AU labels from R. The training procedure for R maximizes the probability of D making a mistake. By leveraging the adversarial mechanism, the distribution of recognized AUs is closed to AU prior distribution from domain knowledge. Furthermore, the proposed weakly supervised AU recognition can be extended to semi-supervised learning scenarios with partially AU-annotated images. Experimental results on three benchmark databases demonstrate that the proposed method successfully leverages the summarized domain knowledge to weakly supervised AU classifier learning through an adversarial process, and thus achieves state-of-the-art performance.

****

Non-Linear Temporal Subspace Representations for Activity Recognition

Anoop Cherian, Suvrit Sra, Stephen Gould, Richard Hartley; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2197-2206

Representations that can compactly and effectively capture the temporal evolution of semantic content are important to computer vision and machine learning algorithms that operate on multi-variate time-series data. We investigate such representations motivated by the task of human action recognition. Here each data instance is encoded by a multivariate feature (such as via a deep CNN) where action dynamics are characterized by their variations in time. As these features are often non-linear, we propose a novel pooling method, kernelized rank pooling, that represents a given sequence compactly as the pre-image of the parameters of a hyperplane in a reproducing kernel Hilbert space, projections of data onto which captures their temporal order. We develop this idea further and show that such a pooling scheme can be cast as an order-constrained kernelized PCA objective. We then propose to use the parameters of a kernelized low-rank feature subspace as the representation of the sequences. We cast our formulation as an optimization problem on generalized Grassmann manifolds and then solve it efficiently using Riemannian optimization techniques. We present experiments on several action recognition datasets using diverse feature modalities and demonstrate state-of-the-art results.

****

Towards Pose Invariant Face Recognition in the Wild

Jian Zhao, Yu Cheng, Yan Xu, Lin Xiong, Jianshu Li, Fang Zhao, Karlekar Jayashree, Sugiri Pranata, Shengmei Shen, Junliang Xing, Shuicheng Yan, Jiashi Feng; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2207-2216

Pose variation is one key challenge in face recognition. As opposed to current techniques for pose invariant face recognition, which either directly extract pose invariant features for recognition, or first normalize profile face images to frontal pose before feature extraction, we argue that it is more desirable to perform both tasks jointly to allow them to benefit from each other. To this end, we propose a Pose Invariant Model (PIM) for face recognition in the wild, with three distinct novelties. First, PIM is a novel and unified deep architecture, containing a Face Frontalization sub-Net (FFN) and a Discriminative Learning sub-Net (DLN), which are jointly learned from end to end. Second, FFN is a well-designed dual-path Generative Adversarial Network (GAN) which simultaneously perceives global structures and local details, incorporated with an unsupervised cross-domain adversarial training and a "learning to learn" strategy for high-fidelity and identity-preserving frontal view synthesis. Third, DLN is a generic Convolutional Neural Network (CNN) for face recognition with our enforced cross-entropy optimization strategy for learning discriminative yet generalized feature representation. Qualitative and quantitative experiments on both controlled and in-the-wild benchmarks demonstrate the superiority of the proposed model over the state-of-the-arts.

****

Unifying Identification and Context Learning for Person Recognition

Qingqiu Huang, Yu Xiong, Dahua Lin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2217-2225

Despite the great success of face recognition techniques, recognizing persons under unconstrained settings remains challenging. Issues like profile views, unfavorable lighting, and occlusions can cause substantial difficulties. Previous works have attempted to tackle this problem by exploiting the context, e.g. clothes and social relations. While showing promising improvement, they are usually limited in two important aspects, relying on simple heuristics to combine different cues and separating the construction of context from people identities. In this work, we aim to move beyond such limitations and propose a new framework to leverage context for person recognition. In particular, we propose a Region Attention Network, which is learned to adaptively combine visual cues with instance-dependent weights. We also develop a unified formulation, where the social contexts are learned along with the reasoning of people identities. These models substantially improve the robustness when working with the complex contextual relations in unconstrained environments. On two large datasets, PIPA and Cast In Movies (CIM), a new dataset proposed in this work, our method consistently achieves state-of-the-art performance under multiple evaluation policies.

****

Jointly Optimize Data Augmentation and Network Training: Adversarial Data Augmentation in Human Pose Estimation

Xi Peng, Zhiqiang Tang, Fei Yang, Rogerio S. Feris, Dimitris Metaxas; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2226-2234

Random data augmentation is a critical technique to avoid overfitting in training deep models. Yet, data augmentation and network training are often two isolated processes in most settings, yielding to a suboptimal training. Why not jointly optimize the two? We propose adversarial data augmentation to address this limitation. The key idea is to design a generator (e.g. an augmentation network) that competes against a discriminator (e.g. a target network) by generating hard examples online. The generator explores weaknesses of the discriminator, while the discriminator learns from hard augmentations to achieve better performance. A reward/penalty strategy is also proposed for efficient joint training. We investigate human pose estimation and carry out comprehensive ablation studies to validate our method. The results prove that our method can effectively improve state-of-the-art models without additional data effort.

****

Wing Loss for Robust Facial Landmark Localisation With Convolutional Neural Networks

Zhen-Hua Feng, Josef Kittler, Muhammad Awais, Patrik Huber, Xiao-Jun Wu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2235-2245

We present a new loss function, namely Wing loss, for robust facial landmark loc alisation with Convolutional Neural Networks (CNNs). We first compare and analyse different loss functions including L2, L1 and smooth L1. The analysis of these loss functions suggests that, for the training of a CNN-based localisation model, more attention should be paid to small and medium range errors. To this end, we design a piece-wise loss function. The new loss amplifies the impact of errors from the interval (-w, w) by switching from L1 loss to a modified logarithm function. To address the problem of under-representation of samples with large out-of-plane head rotations in the training set, we propose a simple but effective boosting strategy, referred to as pose-based data balancing. In particular, we deal with the data imbalance problem by duplicating the minority training samples and perturbing them by injecting random image rotation, bounding box translation and other data augmentation approaches. Last, the proposed approach is extended to create a two-stage framework for robust facial landmark localisation. The experimental results obtained on AFLW and 300W demonstrate the merits of the Wing loss function, and prove the superiority of the proposed method over the state-of-the-art approaches.

****

Multiple Granularity Group Interaction Prediction

Taiping Yao, Minsi Wang, Bingbing Ni, Huawei Wei, Xiaokang Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2246-2254

Most human activity analysis works (i.e., recognition or■prediction) only focus on a single granularity, i.e., either■modelling global motion based on the coarse level movement such as human trajectories or■forecasting future detailed action based on body parts’ movement such as skeleton motion. In contrast, in this work, we propose a multi-granularity interaction prediction network which integrates■both global motion and detailed local action. Built on a bi- directional LSTM network, the■proposed method possesses■between granularities links which encourage feature sharing as well as cross-feature consistency between both global■and local granularity (e.g., trajectory or local action), and in turn predict long-term global location and local dynamics of each individual. We validate our method on several■public datasets with promising performance.

****

Social GAN: Socially Acceptable Trajectories With Generative Adversarial Networks

Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, Alexandre Alahi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2255-2264

Understanding human motion behavior is critical for autonomous moving platforms (like self-driving cars and social robots) if they are to navigate human-centric environments. This is challenging because human motion is inherently multimodal: given a history of human motion paths, there are many socially plausible ways that people could move in the future. We tackle this problem by combining tools from sequence prediction and generative adversarial networks: a recurrent sequence-to-sequence model observes motion histories and predicts future behavior, using a novel pooling mechanism to aggregate information across people. We predict socially plausible futures by training adversarially against a recurrent discriminator, and encourage diverse predictions with a novel variety loss. Through experiments on several datasets we demonstrate that our approach outperforms prior work in terms of accuracy, variety, collision avoidance, and computational complexity.

****

Deep Group-Shuffling Random Walk for Person Re-Identification

Yantao Shen, Hongsheng Li, Tong Xiao, Shuai Yi, Dapeng Chen, Xiaogang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2265-2274

Person re-identification aims at finding a person of interest in an image gallery by comparing the probe image of this person with all the gallery images. It is generally treated as a retrieval problem, where the affinities between the probe image and gallery images (P2G affinities) are used to rank the retrieved galle ry images. However, most existing methods only consider P2G affinities but ignore the affinities between all the gallery images (G2G affinity). Some frameworks incorporated G2G affinities into the testing process, which is not end-to-end trainable for deep neural networks. In this paper, we propose a novel group-shuffling random walk network for fully utilizing the affinity information between gallery images in both the training and testing processes. The proposed approach aims at end-to-end refining the P2G affinities based on G2G affinity information with a simple yet effective matrix operation, which can be integrated into deep neural networks. Feature grouping and group shuffle are also proposed to apply rich supervisions for learning better person features. The proposed approach outperforms state-of-the-art methods on the Market-1501, CUHK03, and DukeMTMC datasets by large margins, which demonstrate the effectiveness of our approach.

****

Transferable Joint Attribute-Identity Deep Learning for Unsupervised Person Re-Identification

Jingya Wang, Xiatian Zhu, Shaogang Gong, Wei Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2275-2284

Most existing person re-identification (re-id) methods require supervised model learning from a separate large set of pairwise labelled training data for every single camera pair. This significantly limits their scalability and usability in real-world large scale deployments with the need for performing re-id across many camera views. To address this scalability problem, we develop a novel deep learning method for transferring the labelled information of an existing dataset to a new unseen (unlabelled) target domain for person re-id without any supervised learning in the target domain. Specifically, we introduce an Transferable Joint Attribute-Identity Deep Learning (TJ-AIDL) for simultaneously learning an attribute-semantic and identitydiscriminative feature representation space transferrable to any new (unseen) target domain for re-id tasks without the need for collecting new labelled training data from the target domain (i.e. unsupervised learning in the target domain). Extensive comparative evaluations validate the superiority of this new TJ-AIDL model for unsupervised person re-id over a wide range of state-of- the-art methods on four challenging benchmarks including VIPeR, PRID, Market-1501, and DukeMTMC-ReID.

****

Harmonious Attention Network for Person Re-Identification

Wei Li, Xiatian Zhu, Shaogang Gong; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2285-2294

Existing person re-identification (re-id) methods either assume the availability of well-aligned person bounding box images as model input or rely on constrained attention selection mechanisms to calibrate misaligned images. They are therefore sub-optimal for re-id matching in arbitrarily aligned person images potentially with large human pose variations and unconstrained auto-detection errors. In this work, we show the advantages of jointly learning attention selection and feature representation in a Convolutional Neural Network (CNN) by maximising the complementary information of different levels of visual attention subject to re-id discriminative learning constraints. Specifically, we formulate a novel Harmonious Attention CNN (HA-CNN) model for joint learning of soft pixel attention and hard regional attention along with simultaneous optimisation of feature representations, dedicated to optimise person re-id in uncontrolled (misaligned) images. Extensive comparative evaluations validate the superiority of this new HACNN model for person re-id over a wide variety of state-of-the-art methods on three large-scale benchmarks including CUHK03, Market-1501, and DukeMTMC-ReID.

****

Real-Time Rotation-Invariant Face Detection With Progressive Calibration Networks

Xuepeng Shi, Shiguang Shan, Meina Kan, Shuzhe Wu, Xilin Chen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2295-2303

Rotation-invariant face detection, i.e. detecting faces with arbitrary rotation-in-plane (RIP) angles, is widely required in unconstrained applications but stil l remains as a challenging task, due to the large variations of face appearances. Most existing methods compromise with speed or accuracy to handle the large RIP variations. To address this problem more efficiently, we propose Progressive Calibration Networks (PCN) to perform rotation-invariant face detection in a coarse-to-fine manner. PCN consists of three stages, each of which not only distinguishes the faces from non-faces, but also calibrates the RIP orientation of each face candidate to upright progressively. By dividing the calibration process into several progressive steps and only predicting coarse orientations in early stages, PCN can achieve precise and fast calibration. By performing binary classification of face vs. non-face with gradually decreasing RIP ranges, PCN can accurately detect faces with full 360^{circ} RIP angles. Such designs lead to a real-time rotation-invariant face detector. The experiments on multi-oriented FDDB and a challenging subset of WIDER FACE containing rotated faces in the wild show that our PCN achieves quite promising performance.

****

Deep Regression Forests for Age Estimation

Wei Shen, Yilu Guo, Yan Wang, Kai Zhao, Bo Wang, Alan L. Yuille; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2304-2313

Age estimation from facial images is typically cast as a nonlinear regression problem. The main challenge of this problem is the facial feature space w.r.t. ages is inhomogeneous, due to the large variation in facial appearance across different persons of the same age and the non-stationary property of aging patterns. In this paper, we propose Deep Regression Forests (DRFs), an end-to-end model, for age estimation. DRFs connect the split nodes to a fully connected layer of a convolutional neural network (CNN) and deal with inhomogeneous data by jointly learning input-dependant data partitions at the split nodes and data abstractions at the leaf nodes. This joint learning follows an alternating strategy: First, by fixing the leaf nodes, the split nodes as well as the CNN parameters are optimized by Back-propagation; Then, by fixing the split nodes, the leaf nodes are optimized by iterating a step-size free update rule derived from Variational Bounding. We verify the proposed DRFs on three standard age estimation benchmarks and achieve state-of-the-art results on all of them.

****

Weakly-Supervised Deep Convolutional Neural Network Learning for Facial Action Unit Intensity Estimation

Yong Zhang, Weiming Dong, Bao-Gang Hu, Qiang Ji; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2314-2323

Facial action unit (AU) intensity estimation plays an important role in affective computing and human-computer interaction. Recent works have introduced deep neural networks for AU intensity estimation, but they require a large amount of intensity annotations. AU annotation needs strong domain expertise and it is expensive to construct a large database to learn deep models. We propose a novel knowledge-based semi-supervised deep convolutional neural network for AU intensity estimation with extremely limited AU annotations. Only the intensity annotations of peak and valley frames in training sequences are needed. To provide additional supervision for model learning, we exploit naturally existing constraints on AUs, including relative appearance similarity, temporal intensity ordering, facial symmetry, and contrastive appearance difference. Experimental evaluations are performed on two public benchmark databases. With around 2% of intensity annotations in FERA 2015 and around 1% in DISFA for training, our method can achieve comparable or even better performance than the state-of-the-art methods which use 100% of intensity annotations in the training set.

****

Memory Based Online Learning of Deep Representations From Video Streams

Federico Pernici, Federico Bartoli, Matteo Bruni, Alberto Del Bimbo; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2324-2334

We present a novel online unsupervised method for face identity learning from video streams. The method exploits deep face descriptors together with a memory ba sed learning mechanism that takes advantage of the temporal coherence of visual data. Specifically, we introduce a discriminative descriptor matching solution based on Reverse Nearest Neighbour and a forgetting strategy that detect redundant descriptors and discard them appropriately while time progresses. It is shown that the proposed learning procedure is asymptotically stable and can be effectively used in relevant applications like multiple face identification and tracking from unconstrained video streams. Experimental results show that the proposed method achieves comparable results in the task of multiple face tracking and better performance in face identification with offline approaches exploiting future information. Code will be publicly available.

****

Efficient and Deep Person Re-Identification Using Multi-Level Similarity

Yiluan Guo, Ngai-Man Cheung; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2335-2344

Person Re-Identification (ReID) requires comparing two images of person captured under different conditions. Existing work based on neural networks often computes the similarity of feature maps from one single convolutional layer. In this work, we propose an efficient, end-to-end fully convolutional Siamese network that computes the similarities at multiple levels. We demonstrate that multi-level similarity can improve the accuracy considerably using low-complexity network structures in ReID problem. Specifically, first, we use several convolutional layers to extract the features of two input images. Then, we propose Convolution Similarity Network to compute the similarity score maps for the inputs. We use spatial transformer networks (STNs) to determine spatial attention. We propose to apply efficient depth-wise convolution to compute the similarity. The proposed Convolution Similarity Networks can be inserted into different convolutional layers to extract visual similarities at different levels. Furthermore, we use an improved ranking loss to further improve the performance. Our work is the first to propose to compute visual similarities at low, middle and high levels for ReID. With extensive experiments and analysis, we demonstrate that our system, compact yet effective, can achieve competitive results with much smaller model size and computational complexity.

****

Multi-Level Fusion Based 3D Object Detection From Monocular Images

Bin Xu, Zhenzhong Chen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2345-2353

In this paper, we present an end-to-end deep learning based framework for 3D object detection from a single monocular image. A deep convolutional neural network is introduced for simultaneous 2D and 3D object detection. First, 2D region proposals are generated through a region proposal network. Then the shared features are learned within the proposals to predict the class probability, 2D bounding box, orientation, dimension, and 3D location. We adopt a stand-alone module to predict the disparity and extract features from the computed point cloud. Thus features from the original image and the point cloud will be fused in different levels for accurate 3D localization. The estimated disparity is also used for front view feature encoding to enhance the input image,regarded as an input-fusionprocess. The proposed algorithm can directly output both 2D and 3D object detection results in an end-to-end fashion with only a single RGB image as the input. The experimental results on the challenging KITTI benchmark demonstrate that our algorithm significantly outperforms the state-of-the-art methods with only monocular images.

****

A Perceptual Measure for Deep Single Image Camera Calibration

Yannick Hold-Geoffroy, Kalyan Sunkavalli, Jonathan Eisenmann, Matthew Fisher, Emiliano Gambaretto, Sunil Hadap, Jean-François Lalonde; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2354-2363

Most current single image camera calibration methods rely on specific image features or user input, and cannot be applied to natural images captured in uncontrolled settings. We propose inferring directly camera calibration parameters from a single image using a deep convolutional neural network. This network is traine d using automatically generated samples from a large-scale panorama dataset, and considerably outperforms other methods, including recent deep learning-based approaches, in terms of standard L2 error. However, we argue that in many cases it is more important to consider how humans perceive errors in camera estimation. To this end, we conduct a large-scale human perception study where we ask users to judge the realism of 3D objects composited with and without ground truth camera calibration. Based on this study, we develop a new perceptual measure for camera calibration, and demonstrate that our deep calibration network outperforms other methods on this measure. Finally, we demonstrate the use of our calibration network for a number of applications including virtual object insertion, image retrieval and compositing.

****

Learning to Generate Time-Lapse Videos Using Multi-Stage Dynamic Generative Adversarial Networks

Wei Xiong, Wenhan Luo, Lin Ma, Wei Liu, Jiebo Luo; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2364-2373

Taking a photo outside, can we predict the immediate future, e.g., how would the cloud move in the sky? We address this problem by presenting a generative adversarial network (GAN) based two-stage approach to generating realistic time-lapse videos of high resolution. Given the first frame, our model learns to generate long-term future frames. The first stage generates videos of realistic contents for each frame. The second stage refines the generated video from the first stage by enforcing it to be closer to real videos with regard to motion dynamics. To further encourage vivid motion in the final generated video, Gram matrix is employed to model the motion more precisely. We build a large scale time-lapse dataset, and test our approach on this new dataset. Using our model, we are able to generate realistic videos of up to $128■imes 128$ resolution for 32 frames. Quantitative and qualitative experiment results have demonstrated the superiority of our model over the state-of-the-art models.

****

Document Enhancement Using Visibility Detection

Netanel Kligler, Sagi Katz, Ayellet Tal; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2374-2382

This paper re-visits classical problems in document enhancement. Rather than proposing a new algorithm for a specific problem, we introduce a novel general approach. The key idea is to modify any state- of-the-art algorithm, by providing it with new information (input), improving its own results. Interestingly, this information is based on a solution to a seemingly unrelated problem of visibility detection in R3. We show that a simple representation of an image as a 3D point cloud, gives visibility detection on this cloud a new interpretation. What does it mean for a point to be visible? Although this question has been widely studied within computer vision, it has always been assumed that the point set is a sampling of a real scene. We show that the answer to this question in our context reveals unique and useful information about the image. We demonstrate the benefit of this idea for document binarization and for unshadowing.

****

A Weighted Sparse Sampling and Smoothing Frame Transition Approach for Semantic Fast-Forward First-Person Videos

Michel Silva, Washington Ramos, João Ferreira, Felipe Chamone, Mario Campos, Erickson R. Nascimento; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2383-2392

Thanks to the advances in the technology of low-cost digital cameras and the popularity of the self-recording culture, the amount of visual data on the Internet is going to the opposite side of the available time and patience of the users. Thus, most of the uploaded videos are doomed to be forgotten and unwatched in a computer folder or website. In this work, we address the problem of creating smooth fast-forward videos without losing the relevant content. We present a new adaptive frame selection formulated as a weighted minimum reconstruction problem, which combined with a smoothing frame transition method accelerates first-person videos emphasizing the relevant segments and avoids visual discontinuities. The experiments show that our method is able to fast-forward videos to retain as much relevant information and smoothness as the state-of-the-art techniques in less time. We also present a new 80-hour multimodal (RGB-D, IMU, and GPS) dataset of first-person videos with annotations for recorder profile, frame scene, activities, interaction, and attention.

****

Context Contrasted Feature and Gated Multi-Scale Aggregation for Scene Segmentation

Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, Gang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2393-2402

Scene segmentation is a challenging task as it need label every pixel in the image. It is crucial to exploit discriminative context and aggregate multi-scale features to achieve better segmentation. In this paper, we first propose a novel context contrasted local feature that not only leverages the informative context but also spotlights the local information in contrast to the context. The proposed context contrasted local feature greatly improves the parsing performance, especially for inconspicuous objects and background stuff. Furthermore, we propose a scheme of gated sum to selectively aggregate multi-scale features for each spatial position. The gates in this scheme control the information flow of different scale features. Their values are generated from the testing image by the proposed network learnt from the training data so that they are adaptive not only to the training data, but also to the specific testing image. Without bells and whistles, the proposed approach achieves the state-of-the-arts consistently on the three popular scene segmentation datasets, Pascal Context, SUN-RGBD and COCO Stuff.

****

Deep Layer Aggregation

Fisher Yu, Dequan Wang, Evan Shelhamer, Trevor Darrell; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2403-2412

Visual recognition requires rich representations that span levels from low to high, scales from small to large, and resolutions from fine to coarse. Even with the depth of features in a convolutional network, a layer in isolation is not enough: compounding and aggregating these representations improves inference of what and where. Architectural efforts are exploring many dimensions for network backbones, designing deeper or wider architectures, but how to best aggregate layers and blocks across a network deserves further attention. Although skip connections have been incorporated to combine layers, these connections have been ``shallow'' themselves, and only fuse by simple, one-step operations. We augment standard architectures with deeper aggregation to better fuse information across layers. Our deep layer aggregation structures iteratively and hierarchically merge the feature hierarchy to make networks with better accuracy and fewer parameters. Experiments across architectures and tasks show that deep layer aggregation improves recognition and resolution compared to existing branching and merging schemes.

****

Convolutional Neural Networks With Alternately Updated Clique

Yibo Yang, Zhisheng Zhong, Tiancheng Shen, Zhouchen Lin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2413-2422

Improving information flow in deep networks helps to ease the training difficulties and utilize parameters more efficiently. Here we propose a new convolutional neural network architecture with alternately updated clique (CliqueNet). In contrast to prior networks, there are both forward and backward connections between any two layers in the same block. The layers are constructed as a loop and are updated alternately. The CliqueNet has some unique properties. For each layer, it is both the input and output of any other layer in the same block, so that the information flow among layers is maximized. During propagation, the newly updated layers are concatenated to re-update previously updated layer, and parameters are reused for multiple times. This recurrent feedback structure is able to bring higher level visual information back to refine low-level filters and achieve spatial attention. We analyze the features generated at different stages and observe that using refined features leads to a better result. We adopt a multi-scale feature strategy that effectively avoids the progressive growth of parameters. Experiments on image recognition datasets including CIFAR-10, CIFAR-100, SVHN and ImageNet show that our proposed models achieve the state-of-the-art performance with fewer parameters.

****

Practical Block-Wise Neural Network Architecture Generation

Zhao Zhong, Junjie Yan, Wei Wu, Jing Shao, Cheng-Lin Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2423-2432

Convolutional neural networks have gained a remarkable success in computer vision. However, most usable network architectures are hand-crafted and usually require expertise and elaborate design. In this paper, we provide a block-wise network generation pipeline called BlockQNN which automatically builds high-performance networks using the Q-Learning paradigm with epsilon-greedy exploration strategy. The optimal network block is constructed by the learning agent which is trained sequentially to choose component layers. We stack the block to construct the whole auto-generated network. To accelerate the generation process, we also propose a distributed asynchronous framework and an early stop strategy.The block-wise generation brings unique advantages: (1) it performs competitive results in comparison to the hand-crafted state-of-the-art networks on image classification, additionally, the best network generated by BlockQNN achieves 3.54% top-1 error rate on CIFAR-10 which beats all existing auto-generate networks. (2) in the meanwhile, it offers tremendous reduction of the search space in designing networks which only spends 3 days with 32 GPUs, and (3) moreover, it has strong generalizability that the network built on CIFAR also performs well on a larger-scale ImageNet dataset.

****

xUnit: Learning a Spatial Activation Function for Efficient Image Restoration

Idan Kligvasser, Tamar Rott Shaham, Tomer Michaeli; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2433-2442

In recent years, deep neural networks (DNNs) achieved unprecedented performance in many low-level vision tasks. However, state-of-the-art results are typically achieved by very deep networks, which can reach tens of layers with tens of millions of parameters. To make DNNs implementable on platforms with limited resources, it is necessary to weaken the tradeoff between performance and efficiency. In this paper, we propose a new activation unit, which is particularly suitable for image restoration problems. In contrast to the widespread per-pixel activation units, like ReLUs and sigmoids, our unit implements a learnable nonlinear function with spatial connections. This enables the net to capture much more complex features, thus requiring a significantly smaller number of layers in order to reach the same performance. We illustrate the effectiveness of our units through experiments with state-of-the-art nets for denoising, de-raining, and super resolution, which are already considered to be very small. With our approach, we are able to further reduce these models by nearly 50% without incurring any degradation in performance.

****

Crafting a Toolchain for Image Restoration by Deep Reinforcement Learning

Ke Yu, Chao Dong, Liang Lin, Chen Change Loy; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2443-2452

We investigate a novel approach for image restoration by reinforcement learning. Unlike existing studies that mostly train a single large network for a specialized task, we prepare a toolbox consisting of small-scale convolutional networks of different complexities and specialized in different tasks. Our method, RL-Restore, then learns a policy to select appropriate tools from the toolbox to progressively restore the quality of a corrupted image. We formulate a step-wise reward function proportional to how well the image is restored at each step to learn the action policy. We also devise a joint learning scheme to train the agent and tools for better performance in handling uncertainty. In comparison to conventional human-designed networks, RL-Restore is capable of restoring images corrupted with complex and unknown distortions in a more parameter-efficient manner using the dynamically formed toolchain.

****

Deformation Aware Image Compression

Tamar Rott Shaham, Tomer Michaeli; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2453-2462

Lossy compression algorithms aim to compactly encode images in a way which enables to restore them with minimal error. We show that a key limitation of existing algorithms is that they rely on error measures that are extremely sensitive to geometric deformations (e.g. SSD, SSIM). These force the encoder to invest many bits in describing the exact geometry of every fine detail in the image, which is obviously wasteful, because the human visual system is indifferent to small local translations. Motivated by this observation, we propose a deformation-insensitive error measure that can be easily incorporated into any existing compression scheme. As we show, optimal compression under our criterion involves slightly deforming the input image such that it becomes more "compressible". Surprisingly, while these small deformations are barely noticeable, they enable the CODEC to preserve details that are otherwise completely lost. Our technique uses the CODEC as a "black box", thus allowing simple integration with arbitrary compression methods. Extensive experiments, including user studies, confirm that our approach significantly improves the visual quality of many CODECs. These include JPEG, JPEG~2000, WebP, BPG, and a recent deep-net method.

****

Distributable Consistent Multi-Object Matching

Nan Hu, Qixing Huang, Boris Thibert, Leonidas J. Guibas; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2463-2471

In this paper we propose an optimization-based framework to multiple object matching. The framework takes maps computed between pairs of objects as input, and outputs maps that are consistent among all pairs of objects. The central idea of our approach is to divide the input object collection into overlapping sub-collections and enforce map consistency among each sub-collection. This leads to a distributed formulation, which is scalable to large-scale datasets. We also present an equivalence condition between this decoupled scheme and the original scheme. Experiments on both synthetic and real-world datasets show that our framework is competitive against state-of-the-art multi-object matching techniques.

****

Residual Dense Network for Image Super-Resolution

Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, Yun Fu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2472-2481

In this paper, we propose dense feature fusion (DFF) for image super-resolution (SR). As the same content in different natural images often have various scales and angles of view, jointly leaning hierarchical features is essential for image SR. On the other hand, very deep convolutional neural network (CNN) has recently achieved great success for image SR and offered hierarchical features as well. However, most of deep CNN based SR models neglect to jointly make full use of the hierarchical features. In addition, dense connected layers would allow the network to be deeper, efficient to train, and more powerful. To embrace these observations, in our proposed DFF model, we fully exploit all the meaningful convolutional features in local and global manners. Specifically, we use dense connected convolutional layers to extract abundant local features. We use local feature fusion to adaptively learn more efficient features from preceding and current local features. After fully obtaining dense local features, we use global feature fusion to jointly and adaptively learn global hierarchical features in a holistic way. Extensive experiments on benchmark datasets show that our DFF achieves favorable performance against state-of-the-art methods quantitatively and visually

.

****

Attentive Generative Adversarial Network for Raindrop Removal From a Single Image

Rui Qian, Robby T. Tan, Wenhan Yang, Jiajun Su, Jiaying Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2482-2491

Raindrops adhered to a glass window or camera lens can severely hamper the visibility of a background scene and degrade an image considerably. In this paper, we address the problem by visually removing raindrops, and thus transforming a raindrop degraded image into a clean one. The problem is intractable, since first the regions occluded by raindrops are not given. Second, the information about the background scene of the occluded regions is completely lost for most part. To resolve the problem, we apply an attentive generative network using adversarial training. Our main idea is to inject visual attention into both the generative and discriminative networks. During the training, our visual attention learns about raindrop regions and their surroundings. Hence, by injecting this information, the generative network will pay more attention to the raindrop regions and the surrounding structures, and the discriminative network will be able to assess the local consistency of the restored regions. This injection of visual attention to both generative and discriminative networks is the main contribution of this paper. Our experiments show the effectiveness of our approach, which outperforms the state of the art methods quantitatively and qualitatively.

****

FSRNet: End-to-End Learning Face Super-Resolution With Facial Priors

Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, Jian Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2492-2501

Face Super-Resolution (SR) is a domain-specific superresolution problem. The facial prior knowledge can be leveraged to better super-resolve face images. We present a novel deep end-to-end trainable Face Super-Resolution Network (FSRNet), which makes use of the geometry prior, i.e., facial landmark heatmaps and parsing maps, to superresolve very low-resolution (LR) face images without wellaligned requirement. Specifically, we first construct a coarse SR network to recover a coarse high-resolution (HR) image. Then, the coarse HR image is sent to two branches: a fine SR encoder and a prior information estimation network, which extracts the image features, and estimates landmark heatmaps/parsing maps respectively. Both image features and prior information are sent to a fine SR decoder to recover the HR image. To generate realistic faces, we also propose the Face Super-Resolution Generative Adversarial Network (FSRGAN) to incorporate the adversarial loss into FSRNet. Further, we introduce two related tasks, face alignment and parsing, as the new evaluation metrics for face SR, which address the inconsistency of classic metrics w.r.t. visual perception. Extensive experiments show that FSRNet and FSRGAN significantly outperforms state of the arts for very LR face SR, both quantitatively and qualitatively.

****

Burst Denoising With Kernel Prediction Networks

Ben Mildenhall, Jonathan T. Barron, Jiawen Chen, Dillon Sharlet, Ren Ng, Robert Carroll; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2502-2510

We present a technique for jointly denoising bursts of images taken from a handheld camera. In particular, we propose a convolutional neural network architecture for predicting spatially varying kernels that can both align and denoise frames, a synthetic data generation approach based on a realistic noise formation model, and an optimization guided by an annealed loss function to avoid undesirable local minima. Our model matches or outperforms the state-of-the-art across a wide range of noise levels on both real and synthetic data.

****

Unsupervised Sparse Dirichlet-Net for Hyperspectral Image Super-Resolution

Ying Qu, Hairong Qi, Chiman Kwan; Proceedings of the IEEE Conference on Computer

Vision and Pattern Recognition (CVPR), 2018, pp. 2511-2520

In many computer vision applications, obtaining images of high resolution in both the spatial and spectral domains are equally important. However, due to hardware limitations, one can only expect to acquire images of high resolution in either the spatial or spectral domains. This paper focuses on hyperspectral image super-resolution (HSI-SR), where a hyperspectral image (HSI) with low spatial resolution (LR) but high spectral resolution is fused with a multispectral image (MSI) with high spatial resolution (HR) but low spectral resolution to obtain HR HSI. Existing deep learning-based solutions are all supervised that would need a large training set and the availability of HR HSI, which is unrealistic. Here, we make the first attempt to solving the HSI-SR problem using an unsupervised encoder-decoder architecture that carries the following uniquenesses. First, it is composed of two encoder-decoder networks, coupled through a shared decoder, in order to preserve the rich spectral information from the HSI network. Second, the network encourages the representations from both modalities to follow a sparse Dirichlet distribution which naturally incorporates the two physical constraints of HSI and MSI. Third, the angular difference between representations are minimized in order to reduce the spectral distortion. We refer to the proposed architecture as unsupervised Sparse Dirichlet-Net, or uSDN. Extensive experimental results demonstrate the superior performance of uSDN as compared to the state-of-the-art.

****

Dynamic Scene Deblurring Using Spatially Variant Recurrent Neural Networks

Jiawei Zhang, Jinshan Pan, Jimmy Ren, Yibing Song, Linchao Bao, Rynson W.H. Lau, Ming-Hsuan Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2521-2529

Due to the spatially variant blur caused by camera shake and object motions under different scene depths, deblurring images captured from dynamic scenes is challenging. Although recent works based on deep neural networks have shown great progress on this problem, their models are usually large and computationally expensive. In this paper, we propose a novel spatially variant neural network to address the problem. The proposed network is composed of three deep convolutional neural networks (CNNs) and a recurrent neural network (RNN). RNN is used as a deconvolution operator performed on feature maps extracted from the input image by one of the CNNs. Another CNN is used to learn the weights for the RNN at every location. As a result, the RNN is spatially variant and could implicitly model the deblurring process with spatially variant kernels. The third CNN is used to reconstruct the final deblurred feature maps into restored image. The whole network is end-to-end trainable. Our analysis shows that the proposed network has a large receptive field even with a small model size. Quantitative and qualitative evaluations on public datasets demonstrate that the proposed method performs favorably against state-of-the-art algorithms in terms of accuracy, speed, and model size.

****

SPLATNet: Sparse Lattice Networks for Point Cloud Processing

Hang Su, Varun Jampani, Deqing Sun, Subhransu Maji, Evangelos Kalogerakis, Ming-Hsuan Yang, Jan Kautz; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2530-2539

We present a network architecture for processing point clouds that directly operates on a collection of points represented as a sparse set of samples in a high-dimensional lattice. Naively applying convolutions on this lattice scales poorly, both in terms of memory and computational cost, as the size of the lattice increases. Instead, our network uses sparse bilateral convolutional layers as building blocks. These layers maintain efficiency by using indexing structures to apply convolutions only on occupied parts of the lattice, and allow flexible specifications of the lattice structure enabling hierarchical and spatially-aware feature learning, as well as joint 2D-3D reasoning. Both point-based and image-based representations can be easily incorporated in a network with such layers and the resulting model can be trained in an end-to-end manner. We present results on 3D segmentation tasks where our approach outperforms existing state-of-the-art t echniques.

****

Surface Networks

Ilya Kostrikov, Zhongshi Jiang, Daniele Panozzo, Denis Zorin, Joan Bruna; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2540-2548

We study data-driven representations for three-dimensional triangle meshes, which are one of the prevalent objects used to represent 3D geometry. Recent works have developed models that exploit the intrinsic geometry of manifolds and graphs, namely the Graph Neural Networks (GNNs) and its spectral variants, which learn from the local metric tensor via the Laplacian operator.   Despite offering excellent sample complexity and built-in invariances, intrinsic geometry alone is invariant to isometric deformations, making it unsuitable for many applications. To overcome this limitation, we propose several upgrades to GNNs to leverage extrinsic differential geometry properties of three-dimensional surfaces, increasing its modeling power. In particular, we propose to exploit the Dirac operator, whose spectrum detects principal curvature directions --- this is in stark contrast with the classical Laplace   operator, which directly measures mean curvature. We coin the resulting models emph{Surface Networks (SN)}. We prove that these models define shape representations that are stable to deformation and to discretization, and we demonstrate the efficiency and versatility of SNs on   two challenging tasks: temporal prediction of mesh deformations under non-linear dynamics and generative models using a variational autoencoder framework with encoders/decoders given by SNs.

****

Self-Supervised Multi-Level Face Model Learning for Monocular Reconstruction at Over 250 Hz

Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, Christian Theobalt; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2549-2559

The reconstruction of dense 3D models of face geometry and appearance from a single image is highly challenging and ill-posed. To constrain the problem, many approaches rely on strong priors, such as parametric face models learned from limited 3D scan data. However, prior models restrict generalization of the true diversity in facial geometry, skin reflectance and illumination. To alleviate this problem, we present the first approach that jointly learns 1) a regressor for face shape, expression, reflectance and illumination on the basis of 2) a concurrently learned parametric face model. Our multi-level face model combines the advantage of 3D Morphable Models for regularization with the out-of-space generalization of a learned corrective space. We train end-to-end on in-the-wild images without dense annotations by fusing a convolutional encoder with a differentiable expert-designed renderer and a self-supervised training loss, both defined at multiple detail levels. Our approach compares favorably to the state-of-the-art in terms of reconstruction quality, better generalizes to real world faces, and runs at over 250Hz.

****

CodeSLAM — Learning a Compact, Optimisable Representation for Dense Visual SLAM

Michael Bloesch, Jan Czarnowski, Ronald Clark, Stefan Leutenegger, Andrew J. Davison; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2560-2568

The representation of geometry in real-time 3D perception systems continues to be a critical research issue. Dense maps capture complete surface shape and can be augmented with semantic labels, but their high dimensionality makes them computationally costly to store and process, and unsuitable for rigorous probabilistic inference. Sparse feature-based representations avoid these problems, but capture only partial scene information and are mainly useful for localisation only. We present a new compact but dense representation of scene geometry which is conditioned on the intensity data from a single image and generated from a code consisting of a small number of parameters. We are inspired by work both on learned depth from images, and auto-encoders. Our approach is suitable for use in a ke yframe-based monocular dense SLAM system: While each keyframe with a code can produce a depth map, the code can be optimised efficiently jointly with pose variables and together with the codes of overlapping keyframes to attain global consistency. Conditioning the depth map on the image allows the code to only represent aspects of the local geometry which cannot directly be predicted from the image. We explain how to learn our code representation, and demonstrate its advantageous properties in monocular SLAM.

****

SGPN: Similarity Group Proposal Network for 3D Point Cloud Instance Segmentation

Weiyue Wang, Ronald Yu, Qiangui Huang, Ulrich Neumann; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2569-2578

We introduce Similarity Group Proposal Network (SGPN), a simple and intuitive deep learning framework for 3D object instance segmentation on point clouds. SGPN uses a single network to predict point grouping proposals and a corresponding semantic class for each proposal, from which we can directly extract instance segmentation results. Important to the effectiveness of SGPN is its novel representation of 3D instance segmentation results in the form of a similarity matrix that indicates the similarity between each pair of points in embedded feature space, thus producing an accurate grouping proposal for each point. To the best of our knowledge, SGPN is the first framework to learn 3D instance-aware semantic segmentation on point clouds. Experimental results on various 3D scenes show the effectiveness of our method on 3D instance segmentation, and we also evaluate the capability of SGPN to improve 3D object detection and semantic segmentation results. We also demonstrate its flexibility by seamlessly incorporating 2D CNN features into the framework to boost performance.

****

PlaneNet: Piece-Wise Planar Reconstruction From a Single RGB Image

Chen Liu, Jimei Yang, Duygu Ceylan, Ersin Yumer, Yasutaka Furukawa; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2579-2588

This paper proposes a deep neural network (DNN) for piece-wise planar depthmap reconstruction from a single RGB image. While DNNs have brought remarkable progress to single-image pixel-wise depth prediction, piece-wise planar depthmap reconstruction requires a structured geometry representation, and has been a difficult task to master even for DNNs. The proposed end-to-end DNN learns to directly infer a set of plane parameters and corresponding plane segmentation masks from a single RGB image. We have generated more than 50,000 piece-wise planar depth maps for training and testing from ScanNet, a large-scale indoor capture database. Our qualitative and quantitative evaluations demonstrate that the proposed approach outperforms baseline methods in terms of both plane segmentation and depth estimation accuracy. To the best of our knowledge, this paper presents the first end-to-end neural architecture for piece-wise planar reconstruction from a single RGB image.

****

Deep Parametric Continuous Convolutional Neural Networks

Shenlong Wang, Simon Suo, Wei-Chiu Ma, Andrei Pokrovsky, Raquel Urtasun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2589-2597

Standard convolutional neural networks assume a grid structured input is available and exploit discrete convolutions as their fundamental building blocks. This limits their applicability to many real-world applications. In this paper we propose Parametric Continuous Convolution, a new learnable operator that operates over non-grid structured data. The key idea is to exploit parameterized kernel functions that span the full continuous vector space. This generalization allows us to learn over arbitrary data structures as long as their support relationship is computable. Our experiments show significant improvement over the state-of-the-art in point cloud segmentation of indoor and outdoor scenes, and lidar motion estimation of driving scenes.

****

FeaStNet: Feature-Steered Graph Convolutions for 3D Shape Analysis

Nitika Verma, Edmond Boyer, Jakob Verbeek; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2598-2606

Convolutional neural networks (CNNs) have massively impacted visual recognition in 2D images, and are now ubiquitous in state-of-the-art approaches. CNNs do not easily extend, however, to data that are not represented by regular grids, such as 3D shape meshes or other graph-structured data, to which traditional local convolution operators do not directly apply. To address this problem, we propose a novel graph-convolution operator to establish correspondences between filter weights and graph neighborhoods with arbitrary connectivity. The key novelty of our approach is that these correspondences are dynamically computed from features learned by the network, rather than relying on predefined static coordinates over the graph as in previous work. We obtain excellent experimental results that significantly improve over previous state-of-the-art shape correspondence results. This shows that our approach can learn effective shape representations from raw input coordinates, without relying on shape descriptors.

****

Image Collection Pop-Up: 3D Reconstruction and Clustering of Rigid and Non-Rigid Categories

Antonio Agudo, Melcior Pijoan, Francesc Moreno-Noguer; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2607-2615

This paper introduces an approach to simultaneously estimate 3D shape, camera pose, and object and type of deformation clustering, from partial 2D annotations in a multi-instance collection of images. Furthermore, we can indistinctly process rigid and non-rigid categories. This advances existing work, which only addresses the problem for one single object or, if multiple objects are considered, they are assumed to be clustered a priori. To handle this broader version of the problem, we model object deformation using a formulation based on multiple unions of subspaces, able to span from small rigid motion to complex deformations. The parameters of this model are learned via Augmented Lagrange Multipliers, in a completely unsupervised manner that does not require any training data at all. Extensive validation is provided in a wide variety of synthetic and real scenarios, including rigid and non-rigid categories with small and large deformations. In all cases our approach outperforms state-of-the-art in terms of 3D reconstruction accuracy, while also providing clustering results that allow segmenting the images into object instances and their associated type of deformation (or action the object is performing).

****

Geometry-Aware Learning of Maps for Camera Localization

Samarth Brahmbhatt, Jinwei Gu, Kihwan Kim, James Hays, Jan Kautz; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2616-2625

Maps are a key component in image-based camera localization and visual SLAM systems: they are used to establish geometric constraints between images, correct drift in relative pose estimation, and relocalize cameras after lost tracking. The exact definitions of maps, however, are often application-specific and hand-crafted for different scenarios (e.g. 3D landmarks, lines, planes, bags of visual words). We propose to represent maps as a deep neural net called MapNet, which enables learning a data-driven map representation. Unlike prior work on learning maps, MapNet exploits cheap and ubiquitous sensory inputs like visual odometry and GPS in addition to images and fuses them together for camera localization. Geometric constraints expressed by these inputs, which have traditionally been used in bundle adjustment or pose-graph optimization, are formulated as loss terms in MapNet training and also used during inference. In addition to directly improving localization accuracy, this allows us to update the MapNet (i.e., maps) in a self-supervised manner using additional unlabeled video sequences from the scene. We also propose a novel parameterization for camera rotation which is better suited for deep-learning based camera pose regression. Experimental results on both the indoor 7-Scenes and the outdoor Oxford RobotCar datasets show significant improvement over prior work. The MapNet project webpage is https://goo.gl/mRB3Au.

****

Recurrent Slice Networks for 3D Segmentation of Point Clouds

Qiangui Huang, Weiyue Wang, Ulrich Neumann; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2626-2635

Point clouds are an efficient data format for 3D data. However, existing 3D segmentation methods for point clouds either do not model local dependencies or require added computations. This work presents a novel 3D segmentation framework, RSNet, to efficiently model local structures in point clouds. The key component of the RSNet is a lightweight local dependency module. It is a combination of a novel slice pooling layer, Recurrent Neural Network (RNN) layers, and a slice unpooling layer. The slice pooling layer is designed to project features of unordered points onto an ordered sequence of feature vectors so that traditional end-to-end learning algorithms (RNNs) can be applied. The performance of RSNet is validated by comprehensive experiments on the S3DIS, ScanNet, and ShapeNet datasets. In its simplest form, RSNets surpass all previous state-of-the-art methods on these benchmarks. And comparisons against previous state-of-the-art methods demonstrate the efficiency of RSNets.

****

Depth-Based 3D Hand Pose Estimation: From Current Achievements to Future Goals

Shanxin Yuan, Guillermo Garcia-Hernando, Björn Stenger, Gyeongsik Moon, Ju Yong Chang, Kyoung Mu Lee, Pavlo Molchanov, Jan Kautz, Sina Honari, Liuhao Ge, Junsong Yuan, Xinghao Chen, Guijin Wang, Fan Yang, Kai Akiyama, Yang Wu, Qingfu Wan, Meysam Madadi, Sergio Escalera, Shile Li, Dongheui Lee, Iason Oikonomidis, Antonis Argyros, Tae-Kyun Kim; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2636-2645

In this paper, we strive to answer two questions: What is the current state of 3D hand pose estimation from depth images? And, what are the next challenges that need to be tackled? Following the successful Hands In the Million Challenge (HIM2017), we investigate the top 10 state-of-the-art methods on three tasks: single frame 3D pose estimation, 3D hand tracking, and hand pose estimation during object interaction. We analyze the performance of different CNN structures with regard to hand shape, joint visibility, view point and articulation distributions. Our findings include: (1) isolated 3D hand pose estimation achieves low mean errors (10 mm) in the view point range of [70, 120] degrees, but it is far from being solved for extreme view points; (2) 3D volumetric representations outperform 2D CNNs, better capturing the spatial structure of the depth data; (3) Discriminative methods still generalize poorly to unseen hand shapes; (4) While joint occlusions pose a challenge for most methods, explicit modeling of structure constraints can significantly narrow the gap between errors on visible and occluded joints.

****

SobolevFusion: 3D Reconstruction of Scenes Undergoing Free Non-Rigid Motion

Miroslava Slavcheva, Maximilian Baust, Slobodan Ilic; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2646-2655

We present a system that builds 3D models of non-rigidly moving surfaces from scratch in real time using a single RGB-D stream. Our solution is based on the variational level set method, thus it copes with arbitrary geometry, including topological changes. It warps a given truncated signed distance field (TSDF) to a target TSDF via gradient flow. Unlike previous approaches that define the gradient using an L2 inner product, our method relies on gradient flow in Sobolev space. Its favourable regularity properties allow for a more straightforward energy formulation that is faster to compute and that achieves higher geometric detail, mitigating the over-smoothing effects introduced by other regularization schemes. In addition, the coarse-to-fine evolution behaviour of the flow is able to handle larger motions, making few frames sufficient for a high-fidelity reconstruction. Last but not least, our pipeline determines voxel correspondences between partial shapes by matching signatures in a low-dimensional embedding of their Laplacian eigenfunctions, and is thus able to reliably colour the output model. A variety of quantitative and qualitative evaluations demonstrate the advantages of our technique.

****

AdaDepth: Unsupervised Content Congruent Adaptation for Depth Estimation

Jogendra Nath Kundu, Phani Krishna Uppala, Anuj Pahuja, R. Venkatesh Babu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2656-2665

Supervised deep learning methods have shown promising results for the task of monocular depth estimation; but acquiring ground truth is costly, and prone to noise as well as inaccuracies. While synthetic datasets have been used to circumvent above problems, the resultant models do not generalize well to natural scenes due to the inherent domain shift. Recent adversarial approaches for domain adaption have performed well in mitigating the differences between the source and target domains. But these methods are mostly limited to a classification setup and do not scale well for fully-convolutional architectures. In this work, we propose AdaDepth - an unsupervised domain adaptation strategy for the pixel-wise regression task of monocular depth estimation. The proposed approach is devoid of above limitations through a) adversarial learning and b) explicit imposition of content consistency on the adapted target representation. Our unsupervised approach performs competitively with other established approaches on depth estimation tasks and achieves state-of-the-art results in a semi-supervised setting.

****

Learning to Find Good Correspondences

Kwang Moo Yi, Eduard Trulls, Yuki Ono, Vincent Lepetit, Mathieu Salzmann, Pascal Fua; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2666-2674

We develop a deep architecture to learn to find good correspondences for wide-baseline stereo. Given a set of putative sparse matches and the camera intrinsics, we train our network in an end-to-end fashion to label the correspondences as inliers or outliers, while simultaneously using them to recover the relative pose, as encoded by the essential matrix. Our architecture is based on a multi-layer perceptron operating on pixel coordinates rather than directly on the image, and is thus simple and small. We introduce a novel normalization technique, called Context Normalization, which allows us to process each data point separately while embedding global information in it, and also makes the network invariant to the order of the correspondences. Our experiments on multiple challenging datasets demonstrate that our method is able to drastically improve the state of the art with little training data.

****

OATM: Occlusion Aware Template Matching by Consensus Set Maximization

Simon Korman, Mark Milam, Stefano Soatto; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2675-2683

We present a novel approach to template matching that is efficient, can handle partial occlusions, and comes with provable performance guarantees. A key component of the method is a reduction that transforms the problem of searching a nearest neighbor among N high-dimensional vectors, to searching neighbors among two sets of order $sqrt{N}$ vectors, which can be found efficiently using range search techniques. This allows for a quadratic improvement in search complexity, and makes the method scalable in handling large search spaces. The second contribution is a hashing scheme based on consensus set maximization, which allows us to handle occlusions. The resulting scheme can be seen as a randomized hypothesize-and-test algorithm, which is equipped with guarantees regarding the number of iterations required for obtaining an optimal solution with high probability. The predicted matching rates are validated empirically and the algorithm shows a significant improvement over the state-of-the-art in both speed and robustness to occlusions.

****

Deep Learning of Graph Matching

Andrei Zanfir, Cristian Sminchisescu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2684-2693

The problem of graph matching under node and pair-wise constraints is fundamental in areas as diverse as combinatorial optimization, machine learning or compute r vision, where representing both the relations between nodes and their neighborhood structure is essential. We present an end-to-end model that makes it possible to learn all parameters of the graph matching process, including the unary and pairwise node neighborhoods, represented as deep feature extraction hierarchies. The challenge is in the formulation of the different matrix computation layers of the model in a way that enables the consistent, efficient propagation of gradients in the complete pipeline from the loss function, through the combinatorial optimization layer solving the matching problem, and the feature extraction hierarchy. Our computer vision experiments and ablation studies on challenging datasets like PASCAL VOC keypoints, Sintel and CUB show that matching models refined end-to-end are superior to counterparts based on feature hierarchies trained for other problems.

****

Unsupervised Discovery of Object Landmarks as Structural Representations

Yuting Zhang, Yijie Guo, Yixin Jin, Yijun Luo, Zhiyuan He, Honglak Lee; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2694-2703

Deep neural networks can model images with rich latent representations, but they cannot naturally conceptualize structures of object categories in a human-perceptible way. This paper addresses the problem of learning object structures in an image modeling process without supervision. We propose an autoencoding formulation to discover landmarks as explicit structural representations. The encoding module outputs landmark coordinates, whose validity is ensured by constraints that reflect the necessary properties for landmarks. The decoding module takes the landmarks as a part of the learnable input representations in an end-to-end differentiable framework. Our discovered landmarks are semantically meaningful and more predictive of manually annotated landmarks than those discovered by previous methods. The coordinates of our landmarks are also complementary features to pretrained deep-neuralnetwork representations in recognizing visual attributes. In addition, the proposed method naturally creates an unsupervised, perceptible interface to manipulate object shapes and decode images with controllable structures.

****

Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, Dmitry Kalenichenko; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2704-2713

The rising popularity of intelligent mobile devices and the daunting computational cost of deep learning-based visual recognition models call for efficient on-device inference schemes. We propose a quantization scheme along with a co-designed training procedure allowing inference to be carried out using integer-only arithmetic while preserving an end-to-end model accuracy that is close to floating-point inference. Inference using integer-only arithmetic performs better than floating-point arithmetic on typical ARM CPUs and can be implemented on integer-arithmetic-only hardware such as mobile accelerators (e.g. Qualcomm Hexagon). By quantizing both activations and weights as 8-bit integers, we obtain a close to 4x memory footprint reduction compared to 32-bit floating-point representations. Even on MobileNets, a model family known for runtime efficiency, our quantization approach results in an improved tradeoff between latency and accuracy on popular ARM CPUs for ImageNet classification and COCO detection.

****

Lean Multiclass Crowdsourcing

Grant Van Horn, Steve Branson, Scott Loarie, Serge Belongie, Pietro Perona; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2714-2723

We introduce a method for efficiently crowdsourcing multiclass annotations in challenging, real world image datasets. Our method is designed to minimize the number of human annotations that are necessary to achieve a desired level of confidence on class labels. It is based on combining models of worker behavior with co mputer vision. Our method is general: it can handle a large number of classes, worker labels that come from a taxonomy rather than a flat list, and can model the dependence of labels when workers can see a history of previous annotations. Our method may be used as a drop-in replacement for the majority vote algorithms used in online crowdsourcing services that aggregate multiple human annotations into a final consolidated label. In experiments conducted on two real-life applications we find that our method can reduce the number of required annotations by as much as a factor of 5.4 and can reduce the residual annotation error by up to 90% when compared with majority voting. Furthermore, the online risk estimates of the models may be used to sort the annotated collection and minimize subsequent expert review effort.

****

Partial Transfer Learning With Selective Adversarial Networks

Zhangjie Cao, Mingsheng Long, Jianmin Wang, Michael I. Jordan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2724-2732

Adversarial learning has been successfully embedded into deep networks to learn transferable features, which reduce distribution discrepancy between the source and target domains. Existing domain adversarial networks assume fully shared label space across domains. In the presence of big data, there is strong motivation of transferring both classification and representation models from existing large-scale domains to unknown small-scale domains. This paper introduces partial transfer learning, which relaxes the shared label space assumption to that the target label space is only a subspace of the source label space. Previous methods typically match the whole source domain to the target domain, which are prone to negative transfer for the partial transfer problem. We present Selective Adversarial Network (SAN), which simultaneously circumvents negative transfer by selecting out the outlier source classes and promotes positive transfer by maximally matching the data distributions in the shared label space. Experiments demonstrate that our models exceed state-of-the-art results for partial transfer learning tasks on several benchmark datasets.

****

Self-Supervised Feature Learning by Learning to Spot Artifacts

Simon Jenni, Paolo Favaro; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2733-2742

We introduce a novel self-supervised learning method based on adversarial training. Our objective is to train a discriminator network to distinguish real images from images with synthetic artifacts, and then to extract features from its intermediate layers that can be transferred to other data domains and tasks. To generate images with artifacts, we pre-train a high-capacity autoencoder and then we use a damage and repair strategy: First, we freeze the autoencoder and damage the output of the encoder by randomly dropping its entries. Second, we augment the decoder with a repair network, and train it in an adversarial manner against the discriminator. The repair network helps generate more realistic images by inpainting the dropped feature entries. To make the discriminator focus on the artifacts, we also make it predict what entries in the feature were dropped. We demonstrate experimentally that features learned by creating and spotting artifacts achieve state of the art performance in several benchmarks.

****

LDMNet: Low Dimensional Manifold Regularized Neural Networks

Wei Zhu, Qiang Qiu, Jiaji Huang, Robert Calderbank, Guillermo Sapiro, Ingrid Daubechies; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2743-2751

Deep neural networks have proved very successful on archetypal tasks for which large training sets are available, but when the training data are scarce, their performance suffers from overfitting. Many existing methods of reducing overfitting are data-independent. Data-dependent regularizations are mostly motivated by the observation that data of interest lie close to a manifold, which is typically hard to parametrize explicitly. These methods usually only focus on the geometry of the input data, and do not necessarily encourage the networks to produce g

eometrically meaningful features. To resolve this, we propose the Low-Dimensional- Manifold-regularized neural Network (LDMNet), which incorporates a feature regularization method that focuses on the geometry of both the input data and the output features. In LDMNet, we regularize the network by encouraging the combination of the input data and the output features to sample a collection of low dimensional manifolds, which are searched efficiently without explicit parametrization. To achieve this, we directly use the manifold dimension as a regularization term in a variational functional. The resulting Euler-Lagrange equation is a Laplace-Beltrami equation over a point cloud, which is solved by the point integral method without increasing the computational complexity. In the experiments, we show that LDMNet significantly outperforms widely-used regularizers. Moreover, LDMNet can extract common features of an object imaged via different modalities, which is very useful in real-world applications such as cross-spectral face recognition.
****

CondenseNet: An Efficient DenseNet Using Learned Group Convolutions
Gao Huang, Shichen Liu, Laurens van der Maaten, Kilian Q. Weinberger; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2752-2761
Deep neural networks are increasingly used on mobile devices, where computational resources are limited. In this paper we develop CondenseNet, a novel network architecture with unprecedented efficiency. It combines dense connectivity with a novel module called learned group convolution. The dense connectivity facilitates feature re-use in the network, whereas learned group convolutions remove connections between layers for which this feature re-use is superfluous. At test time, our model can be implemented using standard group convolutions, allowing for efficient computation in practice. Our experiments show that CondenseNets are far more efficient than state-of-the-art compact convolutional networks such as MobileNets and ShuffleNets.
****

Learning Deep Descriptors With Scale-Aware Triplet Networks
Michel Keller, Zetao Chen, Fabiola Maffra, Patrik Schmuck, Margarita Chli; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2762-2770
Research on learning suitable feature descriptors for Computer Vision has recently shifted to deep learning where the biggest challenge lies with the formulation of appropriate loss functions, especially since the descriptors to be learned are not known at training time. While approaches such as Siamese and triplet losses have been applied with success, it is still not well understood what makes a good loss function. In this spirit, this work demonstrates that many commonly used losses suffer from a range of problems. Based on this analysis, we introduce mixed-context losses and scale-aware sampling, two methods that when combined enable networks to learn consistently scaled descriptors for the first time.
****

Decoupled Networks
Weiyang Liu, Zhen Liu, Zhiding Yu, Bo Dai, Rongmei Lin, Yisen Wang, James M. Rehg, Le Song; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2771-2779
Inner product-based convolution has been a central component of convolutional neural networks (CNNs) and the key to learning visual representations. Inspired by the observation that CNN-learned features are naturally decoupled with the norm of features corresponding to the intra-class variation and the angle corresponding to the semantic difference, we propose a generic decoupled learning framework which models the intra-class variation and semantic difference independently. Specifically, we first reparametrize the inner product to a decoupled form and then generalize it to the decoupled convolution operator which serves as the building block of our decoupled networks. We present several effective instances of the decoupled convolution operator. Each decoupled operator is well motivated and has an intuitive geometric interpretation. Based on these decoupled operators, we further propose to directly learn the operator from data. Extensive experime

nts show that such decoupled reparameterization renders significant performance gain with easier convergence and stronger robustness.

****

Deep Adversarial Metric Learning
Yueqi Duan, Wenzhao Zheng, Xudong Lin, Jiwen Lu, Jie Zhou; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2780-2789

Learning an effective distance metric between image pairs plays an important role in visual analysis, where the training procedure largely relies on hard negative samples. However, hard negatives in the training set usually account for the tiny minority, which may fail to fully describe the distribution of negative samples close to the margin. In this paper, we propose a deep adversarial metric learning (DAML) framework to generate synthetic hard negatives from the observed negative samples, which is widely applicable to supervised deep metric learning methods. Different from existing metric learning approaches which simply ignore numerous easy negatives, the proposed DAML exploits them to generate potential hard negatives adversary to the learned metric as complements. We simultaneously train the hard negative generator and feature embedding in an adversarial manner, so that more precise distance metrics can be learned with adequate and targeted synthetic hard negatives. Extensive experimental results on three benchmark datasets including CUB-200-2011, Cars196 and Stanford Online Products show that DAML effectively boosts the performance of existing deep metric learning approaches through adversarial learning.

****

PU-Net: Point Cloud Upsampling Network
Lequan Yu, Xianzhi Li, Chi-Wing Fu, Daniel Cohen-Or, Pheng-Ann Heng; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2790-2799

Learning and analyzing 3D point clouds with deep networks is challenging due to the sparseness and irregularity of the data. In this paper, we present a data-driven point cloud upsampling technique. The key idea is to learn multi-level features per point and expand the point set via a multi-branch convolution unit implicitly in feature space. The expanded feature is then split to a multitude of features, which are then reconstructed to an upsampled point set. Our network is applied at a patch-level, with a joint loss function that encourages the upsampled points to remain on the underlying surface with a uniform distribution. We conduct various experiments using synthesis and scan data to evaluate our method and demonstrate its superiority over some baseline methods and an optimization-based method. Results show that our upsampled points have better uniformity and are located closer to the underlying surfaces.

****

Real-Time Monocular Depth Estimation Using Synthetic Data With Domain Adaptation via Image Style Transfer
Amir Atapour-Abarghouei, Toby P. Breckon; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2800-2810

Monocular depth estimation using learning-based approaches has become promising in recent years. However, most monocular depth estimators either need to rely on large quantities of ground truth depth data, which is extremely expensive and difficult to obtain, or predict disparity as an intermediary step using a secondary supervisory signal leading to blurring and other artefacts. Training a depth estimation model using pixel-perfect synthetic data can resolve most of these issues but introduces the problem of domain bias. This is the inability to apply a model trained on synthetic data to real-world scenarios. With advances in image style transfer and its connections with domain adaptation (Maximum Mean Discrepancy), we take advantage of style transfer and adversarial training to predict pixel perfect depth from a single real-world color image based on training over a large corpus of synthetic environment data. Experimental results indicate the efficacy of our approach compared to contemporary state-of-the-art techniques.

****

Learning for Disparity Estimation Through Feature Constancy

Zhengfa Liang, Yiliu Feng, Yulan Guo, Hengzhu Liu, Wei Chen, Linbo Qiao, Li Zhou, Jianfeng Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2811-2820

Stereo matching algorithms usually consist of four steps, including matching cost calculation, matching cost aggregation, disparity calculation, and disparity refinement. Existing CNN-based methods only adopt CNN to solve parts of the four steps, or use different networks to deal with different steps, making them difficult to obtain the overall optimal solution. In this paper, we propose a network architecture to incorporate all steps of stereo matching. The network consists of three parts. The first part calculates the multi-scale shared features. The second part performs matching cost calculation, matching cost aggregation and disparity calculation to estimate the initial disparity using shared features. The initial disparity and the shared features are used to calculate the feature constancy that measures correctness of the correspondence between two input images. The initial disparity and the feature constancy are then fed to a sub-network to refine the initial disparity. The proposed method has been evaluated on the Scene Flow and KITTI datasets. It achieves the state-of-the-art performance on the KITTI 2012 and KITTI 2015 benchmarks while maintaining a very fast running time. Source code is available at http://github.com/leonzfa/iResNet.

****

DeepMVS: Learning Multi-View Stereopsis

Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, Jia-Bin Huang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2821-2830

We present DeepMVS, a deep convolutional neural network (ConvNet) for multi-view stereo reconstruction. Taking an arbitrary number of posed images as input, we first produce a set of plane-sweep volumes and use the proposed DeepMVS network to predict high-quality disparity maps. The key contributions that enable these results are (1) supervised pretraining on a photorealistic synthetic dataset, (2) an effective method for aggregating information across a set of unordered images, and (3) integrating multi-layer feature activations from the pre-trained VGG-19 network. We validate the efficacy of DeepMVS using the ETH3D Benchmark. Our results show that DeepMVS compares favorably against state-of-the-art conventional MVS algorithms and other ConvNet based methods, particularly for near-textureless regions and thin structures.

****

Self-Calibrating Polarising Radiometric Calibration

Daniel Teo, Boxin Shi, Yinqiang Zheng, Sai-Kit Yeung; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2831-2839

We present a self-calibrating polarising radiometric calibration method. From a set of images taken from a single viewpoint under different unknown polarising angles, we recover the inverse camera response function and the polarising angles relative to the first angle. The problem is solved in an integrated manner, recovering both of the unknowns simultaneously. The method exploits the fact that the intensity of polarised light should vary sinusoidally as the polarising filter is rotated, provided that the response is linear. It offers the first solution to demonstrate the possibility of radiometric calibration through polarisation. We evaluate the accuracy of our proposed method using synthetic data and real world objects captured using different cameras. The self-calibrated results were found to be comparable with those from multiple exposure sequence.

****

Coding Kendall's Shape Trajectories for 3D Action Recognition

Amor Ben Tanfous, Hassen Drira, Boulbaba Ben Amor; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2840-2849

Suitable shape representations as well as their temporal evolution, termed trajectories, often lie to non-linear manifolds. This puts an additional constraint (i.e., non-linearity) in using conventional machine learning techniques for the purpose of classification, event detection, prediction, etc. This paper accommodates the well-known Sparse Coding and Dictionary Learning to the Kendall's shape space and illustrates effective coding of 3D skeletal sequences for action recog nition. Grounding on the Riemannian geometry of the shape space, an intrinsic sparse coding and dictionary learning formulation is proposed for static skeletal shapes to overcome the inherent non-linearity of the manifold. As a main result, initial trajectories give rise to sparse code functions with suitable computational properties, including sparsity and vector space representation. To achieve action recognition, two different classification schemes were adopted. A bi-directional LSTM is directly performed on sparse code functions, while a linear SVM is applied after representing sparse code functions using Fourier temporal pyramid. Experiments conducted on three publicly available datasets show the superiority of the proposed approach compared to existing Riemannian representations and its competitiveness with respect to other recently-proposed approaches. When the benefits of invariance are maintained from the Kendall's shape representation, our approach not only overcomes the problem of non-linearity but also yields to discriminative sparse code functions.

****

Efficient, Sparse Representation of Manifold Distance Matrices for Classical Scaling

Javier S. Turek, Alexander G. Huth; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2850-2858

Geodesic distance matrices can reveal shape properties that are largely invariant to non-rigid deformations, and thus are often used to analyze and represent 3-D shapes. However, these matrices grow quadratically with the number of points. Thus for large point sets it is common to use a low-rank approximation to the distance matrix, which fits in memory and can be efficiently analyzed using methods such as multidimensional scaling (MDS). In this paper we present a novel sparse method for efficiently representing geodesic distance matrices using biharmonic interpolation. This method exploits knowledge of the data manifold to learn a sparse interpolation operator that approximates distances using a subset of points. We show that our method is 2x faster and uses 20x less memory than current leading methods for solving MDS on large point sets, with similar quality. This enables analyses of large point sets that were previously infeasible.

****

Motion Segmentation by Exploiting Complementary Geometric Models

Xun Xu, Loong Fah Cheong, Zhuwen Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2859-2867

Many real-world sequences cannot be conveniently categorized as general or degenerate; in such cases, imposing a false dichotomy in using the fundamental matrix or homography model for motion segmentation would lead to difficulty. Even when we are confronted with a general scene-motion, the fundamental matrix approach as a model for motion segmentation still suffers from several defects, which we discuss in this paper. The full potential of the fundamental matrix approach could only be realized if we judiciously harness information from the simpler homography model. From these considerations, we propose a multi-view spectral clustering framework that synergistically combines multiple models together. We show that the performance can be substantially improved in this way. We perform extensive testing on existing motion segmentation datasets, achieving state-of-the-art performance on all of them; we also put forth a more realistic and challenging dataset adapted from the KITTI benchmark, containing real-world effects such as strong perspectives and strong forward translations not seen in the traditional datasets.

****

Estimation of Camera Locations in Highly Corrupted Scenarios: All About That Base, No Shape Trouble

Yunpeng Shi, Gilad Lerman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2868-2876

We propose a strategy for improving camera location estimation in structure from motion. Our setting assumes highly corrupted pairwise directions (i.e., normalized relative location vectors), so there is a clear room for improving current state-of-the-art solutions for this problem. Our strategy identifies severely corrupted pairwise directions by using a geometric consistency condition. It then s elects a cleaner set of pairwise directions as a preprocessing step for common solvers. We theoretically guarantee the successful performance of a basic version of our strategy under a synthetic corruption model. Numerical results on artificial and real data demonstrate the significant improvement obtained by our strategy.

****

4D Human Body Correspondences From Panoramic Depth Maps

Zhong Li, Minye Wu, Wangyiteng Zhou, Jingyi Yu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2877-2886

The availability of affordable 3D full body reconstruction systems has given rise to free-viewpoint video (FVV) of human avatars. Most existing solutions produce temporally uncorrelated point clouds or meshes with unknown point/vertex correspondences. Individually compressing each frame is ineffective and still yields to ultra-large data sizes. We present an end-to-end deep learning scheme to establish dense shape correspondences and subsequently compress the data. Our approach uses sparse set of "panoramic" depth maps or PDMs, each emulating an inward-viewing concentric mosaics (CM). We then develop a learning-based technique to learn pixel-wise feature descriptors on PDMs. The results are fed into an autoencoder-based network for compression. Comprehensive experiments demonstrate our solution is robust and effective on both public and our newly captured datasets.

****

Reconstructing Thin Structures of Manifold Surfaces by Integrating Spatial Curves

Shiwei Li, Yao Yao, Tian Fang, Long Quan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2887-2896

The manifold surface reconstruction in multi-view stereo often fails in retaining thin structures due to incomplete and noisy reconstructed point clouds. In this paper, we address this problem by leveraging spatial curves. The curve representation in nature is advantageous in modeling thin and elongated structures, implying topology and connectivity information of the underlying geometry, which exactly compensates the weakness of scattered point clouds. We present a novel surface reconstruction method using both curves and point clouds. First, we propose a 3D curve reconstruction algorithm based on the initialize-optimize-expand strategy. Then, tetrahedra are constructed from points and curves, where the volumes of thin structures are robustly preserved by the Curve-conformed Delaunay Refinement. Finally, the mesh surface is extracted from tetrahedra by a graph optimization. The method has been intensively evaluated on both synthetic and real-world datasets, showing significant improvements over state-of-the-art methods.

****

Multi-View Consistency as Supervisory Signal for Learning Shape and Pose Prediction

Shubham Tulsiani, Alexei A. Efros, Jitendra Malik; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2897-2905

We present a framework for learning single-view shape and pose prediction without using direct supervision for either. Our approach allows leveraging multi-view observations from unknown poses as supervisory signal during training. Our proposed training setup enforces geometric consistency between the independently predicted shape and pose from two views of the same instance. We consequently learn to predict shape in an emergent canonical (view-agnostic) frame along with a corresponding pose predictor. We show empirical and qualitative results using the ShapeNet dataset and observe encouragingly competitive performance to previous techniques which rely on stronger forms of supervision. We also demonstrate the applicability of our framework in a realistic setting which is beyond the scope of existing techniques: using a training dataset comprised of online product images where the underlying shape and pose are unknown.

****

Probabilistic Plant Modeling via Multi-View Image-to-Image Translation

Takahiro Isokane, Fumio Okura, Ayaka Ide, Yasuyuki Matsushita, Yasushi Yagi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2906-2915

This paper describes a method for inferring three-dimensional (3D) plant branch structures that are hidden under leaves from multi-view observations. Unlike previous geometric approaches that heavily rely on the visibility of the branches or use parametric branching models, our method makes statistical inferences of branch structures in a probabilistic framework. By inferring the probability of branch existence using a Bayesian extension of image-to-image translation applied to each of multi-view images, our method generates a probabilistic plant 3D model, which represents the 3D branching pattern that cannot be directly observed. Experiments demonstrate the usefulness of the proposed approach in generating convincing branch structures in comparison to prior approaches.

****

Deep Marching Cubes: Learning Explicit Surface Representations

Yiyi Liao, Simon Donné, Andreas Geiger; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2916-2925

Existing learning based solutions to 3D surface prediction cannot be trained end-to-end as they operate on intermediate representations (e.g., TSDF) from which 3D surface meshes must be extracted in a post-processing step (e.g., via the marching cubes algorithm). In this paper, we investigate the problem of end-to-end 3D surface prediction. We first demonstrate that the marching cubes algorithm is not differentiable and propose an alternative differentiable formulation which we insert as a final layer into a 3D convolutional neural network. We further propose a set of loss functions which allow for training our model with sparse point supervision. Our experiments demonstrate that the model allows for predicting sub-voxel accurate 3D shapes of arbitrary topology. Additionally, it learns to complete shapes and to separate an object's inside from its outside even in the presence of sparse and incomplete ground truth. We investigate the benefits of our approach on the task of inferring shapes from 3D point clouds. Our model is flexible and can be combined with a variety of shape encoder and shape inference techniques.

****

Tags2Parts: Discovering Semantic Regions From Shape Tags

Sanjeev Muralikrishnan, Vladimir G. Kim, Siddhartha Chaudhuri; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2926-2935

We propose a novel method for discovering shape regions that strongly correlate with user-prescribed tags. For example, given a collection of chairs tagged as either "has armrest" or "lacks armrest", our system correctly highlights the armrest regions as the main distinctive parts between the two chair types. To obtain point-wise predictions from shape-wise tags we develop a novel neural network architecture that is trained with tag classification loss, but is designed to rely on segmentation to predict the tag. Our network is inspired by U-Net, but we replicate shallow U structures several times with new skip connections and pooling layers, and call the resulting architecture "WU-Net". We test our method on segmentation benchmarks and show that even with weak supervision of whole shape tags, our method can infer meaningful semantic regions, without ever observing shape segmentations. Further, once trained, the model can process shapes for which the tag is entirely unknown. As a bonus, our architecture is directly operational under full supervision and performs strongly on standard benchmarks. We validate our method through experiments with many variant architectures and prior baselines, and demonstrate several applications.

****

Uncalibrated Photometric Stereo Under Natural Illumination

Zhipeng Mo, Boxin Shi, Feng Lu, Sai-Kit Yeung, Yasuyuki Matsushita; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2936-2945

This paper presents a photometric stereo method that works with unknown natural illuminations without any calibration object. To solve this challenging problem, we propose the use of an equivalent directional lighting model for small surface patches consisting of slowly varying normals, and solve each patch up to an arbitrary rotation ambiguity. Our method connects the resulting patches and unifie s the local ambiguities to a global rotation one through angular distance propagation defined over the whole surface. After applying the integrability constraint, our final solution contains only a binary ambiguity, which could be easily removed. Experiments using both synthetic and real-world datasets show our method provides even comparable results to calibrated methods

****

Robust Depth Estimation From Auto Bracketed Images

Sunghoon Im, Hae-Gon Jeon, In So Kweon; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2946-2954

As demand for advanced photographic applications on hand-held devices grows, these electronics require the capture of high quality depth. However, under low-light conditions, most devices still suffer from low imaging quality and inaccurate depth acquisition. To address the problem, we present a robust depth estimation method from a short burst shot with varied intensity (i.e., Auto Bracketing) or strong noise (i.e., High ISO). We introduce a geometric transformation between flow and depth tailored for burst images, enabling our learning-based multi-view stereo matching to be performed effectively. We then describe our depth estimation pipeline that incorporates the geometric transformation into our residual-flow network. It allows our framework to produce an accurate depth map even with a bracketed image sequence. We demonstrate that our method outperforms state-of-the-art methods for various datasets captured by a smartphone and a DSLR camera. Moreover, we show that the estimated depth is applicable for image quality enhancement and photographic editing.

****

Free Supervision From Video Games

Philipp Krähenbühl; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2955-2964

Deep networks are extremely hungry for data. They devour hundreds of thousands of labeled images to learn robust and semantically meaningful feature representations. Current networks are so data hungry that collecting labeled data has become as important as designing the networks themselves. Unfortunately, manual data collection is both expensive and time consuming. We present an alternative, and show how ground truth labels for many vision tasks are easily extracted from video games in real time as we play them. We interface the popular Microsoft DirectX rendering API, and inject specialized rendering code into the game as it is running. This code produces ground truth labels for instance segmentation, semantic labeling, depth estimation, optical flow, intrinsic image decomposition, and instance tracking. Instead of labeling images, a researcher now simply plays video games all day long. Our method is general and works on a wide range of video games. We collected a dataset of 220k training images, and 60k test images across 3 video games, and evaluate state of the art optical flow, depth estimation and intrinsic image decomposition algorithms. Our video game data is visually closer to real world images, than other synthetic dataset.

****

Planar Shape Detection at Structural Scales

Hao Fang, Florent Lafarge, Mathieu Desbrun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2965-2973

Interpreting 3D data such as point clouds or surface meshes depends heavily on the scale of observation. Yet, existing algorithms for shape detection rely on trial-and-error parameter tunings to output configurations representative of a structural scale. We present a framework to automatically extract a set of representations that capture the shape and structure of man-made objects at different key abstraction levels. A shape-collapsing process first generates a fine-to-coarse sequence of shape representations by exploiting local planarity. This sequence is then analyzed to identify significant geometric variations between successive representations through a supervised energy minimization. Our framework is flexible enough to learn how to detect both existing structural formalisms such as the CityGML Levels Of Details, and expert-specified levels of abstraction. Experiments on different input data and classes of man-made objects, as well as comparisons with existing shape detection methods, illustrate the strengths of our ap proach in terms of efficiency and flexibility.

****

Pix3D: Dataset and Methods for Single-Image 3D Shape Modeling

Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B. Tenenbaum, William T. Freeman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2974-2983

We study 3D shape modeling from a single image and make contributions to it in three aspects. First, we present Pix3D, a large-scale benchmark of diverse image-shape pairs with pixel-level 2D-3D alignment. Pix3D has wide applications in shape-related tasks including reconstruction, retrieval, viewpoint estimation, etc. Building such a large-scale dataset, however, is highly challenging; existing datasets either contain only synthetic data, or lack precise alignment between 2D images and 3D shapes, or only have a small number of images. Second, we calibrate the evaluation criteria for 3D shape reconstruction through behavioral studies, and use them to objectively and systematically benchmark cutting-edge reconstruction algorithms on Pix3D. Third, we design a novel model that simultaneously performs 3D reconstruction and pose estimation; our multi-task learning approach achieves state-of-the-art performance on both tasks.

****

Camera Pose Estimation With Unknown Principal Point

Viktor Larsson, Zuzana Kukelova, Yinqiang Zheng; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2984-2992

To estimate the 6-DoF extrinsic pose of a pinhole camera with partially unknown intrinsic parameters is a critical sub-problem in structure-from-motion and camera localization. In most of existing camera pose estimation solvers, the principal point is assumed to be in the image center. Unfortunately, this assumption is not always true, especially for asymmetrically cropped images. In this paper, we develop the first exactly minimal solver for the case of unknown principal point and focal length by using four and a half point correspondences (P4.5Pfuv). We also present an extremely fast solver for the case of unknown aspect ratio (P5Pfuva). The new solvers outperform the previous state-of-the-art in terms of stability and speed. Finally, we explore the extremely challenging case of both unknown principal point and radial distortion, and develop the first practical non-minimal solver by using seven point correspondences (P7Pfruv). Experimental results on both simulated data and real Internet images demonstrate the usefulness of our new solvers.

****

Inverse Composition Discriminative Optimization for Point Cloud Registration

Jayakorn Vongkulbhisal, Beñat Irastorza Ugalde, Fernando De la Torre, João P. Costeira; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2993-3001

Rigid Point Cloud Registration (PCReg) refers to the problem of finding the rigid transformation between two sets of point clouds. This problem is particularly important due to the advances in new 3D sensing hardware, and it is challenging because neither the correspondence nor the transformation parameters are known. Traditional local PCReg methods (e.g., ICP) rely on local optimization algorithms, which can get trapped in bad local minima in the presence of noise, outliers, bad initializations, etc. To alleviate these issues, this paper proposes Inverse Composition Discriminative Optimization (ICDO), an extension of Discriminative Optimization (DO), which learns a sequence of update steps from synthetic training data that search the parameter space for an improved solution. Unlike DO, ICDO is object-independent and generalizes even to unseen shapes. We evaluated ICDO on both synthetic and real data, and show that ICDO can match the speed and outperform the accuracy of state-of-the-art PCReg algorithms.

****

SurfConv: Bridging 3D and 2D Convolution for RGBD Images

Hang Chu, Wei-Chiu Ma, Kaustav Kundu, Raquel Urtasun, Sanja Fidler; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3002-3011

The last few years have seen approaches trying to combine the increasing popular

ity of depth sensors and the success of the convolutional neural networks. Using depth as additional channel alongside the RGB input has the scale variance problem present in image convolution based approaches. On the other hand, 3D convolution wastes a large amount of memory on mostly unoccupied 3D space, which consists of only the surface visible to the sensor. Instead, we propose SurfConv, which “slides” compact 2D filters along the visible 3D surface. SurfConv is formulated as a simple depth-aware multi-scale 2D convolution, through a new Data-Driven Depth Discretization (D4) scheme. We demonstrate the effectiveness of our method on indoor and outdoor 3D semantic segmentation datasets. Our method achieves state-of-the-art performance while using less than 30% parameters used by the 3D convolution based approaches.
****

A Fast Resection-Intersection Method for the Known Rotation Problem
Qianggong Zhang, Tat-Jun Chin, Huu Minh Le; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3012-3021
The known rotation problem refers to a special case of structure-from-motion where the absolute orientations of the cameras are known. When formulated as a minimax (l_infty) problem on reprojection errors, the problem is an instance of pseudo-convex programming. Though theoretically tractable, solving the known rotation problem on large-scale data (1,000’s of views, 10,000’s scene points) using existing methods can be very time-consuming. In this paper, we devise a fast algorithm for the known rotation problem. Our approach alternates between pose estimation and triangulation (i.e., resection-intersection) to break the problem into multiple simpler instances of pseudo-convex programming. The key to the vastly superior performance of our method lies in using a novel minimum enclosing ball (MEB) technique for the calculation of updating steps, which obviates the need for convex optimisation routines and greatly reduces memory footprint. We demonstrate the practicality of our method on large-scale problem instances which easily overwhelm current state-of-the-art algorithms (demo program available in supplementary).
****

3D Pose Estimation and 3D Model Retrieval for Objects in the Wild
Alexander Grabner, Peter M. Roth, Vincent Lepetit; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3022-3031
We propose a scalable, efficient and accurate approach to retrieve 3D models for objects in the wild. Our contribution is twofold. We first present a 3D pose estimation approach for object categories which significantly outperforms the state-of-the-art on Pascal3D+. Second, we use the estimated pose as a prior to retrieve 3D models which accurately represent the geometry of objects in RGB images. For this purpose, we render depth images from 3D models under our predicted pose and match learned image descriptors of RGB images against those of rendered depth images using a CNN-based multi-view metric learning approach. In this way, we are the first to report quantitative results for 3D model retrieval on Pascal3D+, where our method chooses the same models as human annotators for 50% of the validation images on average. In addition, we show that our method, which was trained purely on Pascal3D+, retrieves rich and accurate 3D models from ShapeNet given RGB images of objects in the wild.
****

Structure From Recurrent Motion: From Rigidity to Recurrency
Xiu Li, Hongdong Li, Hanbyul Joo, Yebin Liu, Yaser Sheikh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3032-3040
This paper proposes a new method for Non-rigidstructure-from-motion (NRSfM). Departing significantlyfrom the traditional idea of using linear low-order shapemodel for NRSfM, our method exploits the property of shaperecurrence (i.e. many dynamic shapes tend to repeat them-selves in time). We show that recurrency is in fact agen-eralized rigidity. Based on this, we show how to reduceNRSfM problems to rigid ones, provided that the recurrencecondition is satisfied. Given such a reduction, standardrigid-SFM techniques can be applied directly (without anychange) to reconstruct the non-rigid d

ynamic shape. To im-plement this idea as a practical approach, this paper de-velops efficient and reliable algorithm for automatic recur-rence detection, as well as new method for camera viewsclustering via rigidity-check. Experiments on both syntheticsequences and real data demonstrate the effectiveness of theproposed method. Since the method provides novel perspec-tive to look at Structure-from-Motion, we hope it will inspireother new researches in the field.

****

Learning Patch Reconstructability for Accelerating Multi-View Stereo

Alex Poms, Chenglei Wu, Shoou-I Yu, Yaser Sheikh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3041-3050

We present an approach to accelerate multi-view stereo (MVS) by prioritizing computation on image patches that are likely to produce accurate 3D surface reconstructions. Our key insight is that the accuracy of the surface reconstruction from a given image patch can be predicted significantly faster than performing the actual stereo matching. The intuition is that non-specular, fronto-parallel, in-focus patches are more likely to produce accurate surface reconstructions than highly specular, slanted, blurry patches --- and that these properties can be reliably predicted from the image itself. By prioritizing stereo matching on a subset of patches that are highly reconstructable and also cover the 3D surface, we are able to accelerate MVS with minimal reduction in accuracy and completeness. To predict the reconstructability score of an image patch from a single view, we train an image-to-reconstructability neural network: the I2RNet. This reconstructability score enables us to efficiently identify image patches that are likely to provide the most accurate surface estimates before performing stereo matching. We demonstrate that the I2RNet, when trained on the ScanNet dataset, generalizes to the DTU and Tanks and Temples MVS datasets. By using our I2RNet with an existing MVS implementation, we show that our method can achieve more than a 30x speed-up over the baseline with only an minimal loss in completeness.

****

Progressively Complementarity-Aware Fusion Network for RGB-D Salient Object Detection

Hao Chen, Youfu Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3051-3060

How to incorporate cross-modal complementarity sufficiently is the cornerstone question for RGB-D salient object detection. Previous works mainly address this issue by simply concatenating multi-modal features or combining unimodal predictions. In this paper, we answer this question from two perspectives: (1) We argue that if the complementary part can be modelled more explicitly, the cross-modal complement is likely to be better captured. To this end, we design a novel complementarity-aware fusion (CA-Fuse) module when adopting the Convolutional Neural Network (CNN). By introducing cross-modal residual functions and complementarity-aware supervisions in each CA-Fuse module, the problem of learning complementary information from the paired modality is explicitly posed as asymptotically approximating the residual function. (2) Exploring the complement across all the levels. By cascading the CA-Fuse module and adding level-wise supervision from deep to shallow densely, the cross-level complement can be selected and combined progressively. The proposed RGB-D fusion network disambiguates both cross-modal and cross-level fusion processes and enables more sufficient fusion results. The experiments on public datasets show the effectiveness of the proposed CA-Fuse module and the RGB-D salient object detection network.

****

Pixels, Voxels, and Views: A Study of Shape Representations for Single View 3D Object Shape Prediction

Daeyun Shin, Charless C. Fowlkes, Derek Hoiem; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3061-3069

The goal of this paper is to compare surface-based and volumetric 3D object shape representations, as well as viewer-centered and object-centered reference frames for single-view 3D shape prediction. We propose a new algorithm for predicting depth maps from multiple viewpoints, with a single depth or RGB image as input. By modifying the network and the way models are evaluated, we can directly co mpare the merits of voxels vs. surfaces and viewer-centered vs. object-centered for familiar vs. unfamiliar objects, as predicted from RGB or depth images. Among our findings, we show that surface-based methods outperform voxel representations for objects from novel classes and produce higher resolution outputs. We also find that using viewer-centered coordinates is advantageous for novel objects, while object-centered representations are better for more familiar objects. Interestingly, the coordinate frame significantly affects the shape representation learned, with object-centered placing more importance on implicitly recognizing the object category and viewer-centered producing shape representations with less dependence on category recognition.

****

Learning Dual Convolutional Neural Networks for Low-Level Vision

Jinshan Pan, Sifei Liu, Deqing Sun, Jiawei Zhang, Yang Liu, Jimmy Ren, Zechao Li, Jinhui Tang, Huchuan Lu, Yu-Wing Tai, Ming-Hsuan Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3070-3079

In this paper, we propose a general dual convolutional neural network (DualCNN) for low-level vision problems, e.g., super-resolution, edge-preserving filtering, deraining and dehazing. These problems usually involve the estimation of two components of the target signals: structures and details. Motivated by this, our proposed DualCNN consists of two parallel branches, which respectively recovers the structures and details in an end-to-end manner. The recovered structures and details can generate the target signals according to the formation model for each particular application. The DualCNN is a flexible framework for low-level vision tasks and can be easily incorporated with existing CNNs. Experimental results show that the DualCNN can be effectively applied to numerous low-level vision tasks with favorable performance against the state-of-the-art methods.

****

Defocus Blur Detection via Multi-Stream Bottom-Top-Bottom Fully Convolutional Network

Wenda Zhao, Fan Zhao, Dong Wang, Huchuan Lu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3080-3088

Defocus blur detection (DBD) is the separation of infocus and out-of-focus regions in an image. This process has been paid considerable attention because of its remarkable potential applications. Accurate differentiation of homogeneous regions and detection of low-contrast focal regions, as well as suppression of background clutter, are challenges associated with DBD. To address these issues, we propose a multi-stream bottom-top-bottom fully convolutional network (BTBNet), which is the first attempt to develop an end-to-end deep network for DBD. First, we develop a fully convolutional BTBNet to integrate low-level cues and high-level semantic information. Then, considering that the degree of defocus blur is sensitive to scales, we propose multi-stream BTBNets that handle input images with different scales to improve the performance of DBD. Finally, we design a fusion and recursive reconstruction network to recursively refine the preceding blur detection maps. To promote further study and evaluation of the DBD models, we construct a new database of 500 challenging images and their pixel-wise defocus blur annotations. Experimental results on the existing and our new datasets demonstrate that the proposed method achieves significantly better performance than other state-of-the-art algorithms.

****

PiCANet: Learning Pixel-Wise Contextual Attention for Saliency Detection

Nian Liu, Junwei Han, Ming-Hsuan Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3089-3098

Contexts play an important role in the saliency detection task. However, given a context region, not all contextual information is helpful for the final task. In this paper, we propose a novel pixel-wise contextual attention network, i.e., the PiCANet, to learn to selectively attend to informative context locations for each pixel. Specifically, for each pixel, it can generate an attention map in which each attention weight corresponds to the contextual relevance at each context location. An attended contextual feature can then be constructed by selective ly aggregating the contextual information. We formulate the proposed PiCANet in both global and local forms to attend to global and local contexts, respectively. Both models are fully differentiable and can be embedded into CNNs for joint training. We also incorporate the proposed models with the U-Net architecture to detect salient objects. Extensive experiments show that the proposed PiCANets can consistently improve saliency detection performance. The global and local PiCANets facilitate learning global contrast and homogeneousness, respectively. As a result, our saliency model can detect salient objects more accurately and uniformly, thus performing favorably against the state-of-the-art methods.

****

Curve Reconstruction via the Global Statistics of Natural Curves

Ehud Barnea, Ohad Ben-Shahar; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3099-3107

Reconstructing the missing parts of a curve has been the subject of much computational research, with applications in image inpainting, object synthesis, etc. Different approaches for solving that problem are typically based on processes that seek visually pleasing or perceptually plausible completions. In this work we focus on reconstructing the underlying physically likely shape by utilizing the global statistics of natural curves. More specifically, we develop a reconstruction model that seeks the mean physical curve for a given inducer configuration. This simple model is both straightforward to compute and it is receptive to diverse additional information, but it requires enough samples for all curve configurations, a practical requirement that limits its effective utilization. To address this practical issue we explore and exploit statistical geometrical properties of natural curves, and in particular, we show that in many cases the mean curve is scale invariant and often times it is extensible. This, in turn, allows to boost the number of examples and thus the robustness of the statistics and its applicability. The reconstruction results are not only more physically plausible but they also lead to important insights on the reconstruction problem, including an elegant explanation why certain inducer configurations are more likely to yield consistent perceptual completions than others.

****

What Do Deep Networks Like to See?

Sebastian Palacio, Joachim Folz, Jörn Hees, Federico Raue, Damian Borth, Andreas Dengel; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3108-3117

We propose a novel way to measure and understand convolutional neural networks by quantifying the amount of input signal they let in. To do this, an autoencoder (AE) was fine-tuned on gradients from a pre-trained classifier with fixed parameters. We compared the reconstructed samples from AEs that were fine-tuned on a set of image classifiers (AlexNet, VGG16, ResNet-50, and Inception~v3) and found substantial differences. The AE learns which aspects of the input space to preserve and which ones to ignore, based on the information encoded in the backpropagated gradients. Measuring the changes in accuracy when the signal of one classifier is used by a second one, a relation of total order emerges. This order depends directly on each classifier's input signal but it does not correlate with classification accuracy or network size. Further evidence of this phenomenon is provided by measuring the normalized mutual information between original images and auto-encoded reconstructions from different fine-tuned AEs. These findings break new ground in the area of neural network understanding, opening a new way to reason, debug, and interpret their results. We present four concrete examples in the literature where observations can now be explained in terms of the input signal that a model uses.

****

“Zero-Shot” Super-Resolution Using Deep Internal Learning

Assaf Shocher, Nadav Cohen, Michal Irani; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3118-3126

Deep Learning has led to a dramatic leap in Super-Resolution (SR) performance in the past few years. However, being supervised, these SR methods are restricted to specific training data, where the acquisition of the low-resolution (LR) imag es from their high-resolution (HR) counterparts is predetermined (e.g., bicubic downscaling), without any distracting artifacts (e.g., sensor noise, image compression, non-ideal PSF, etc). Real LR images, however, rarely obey these restrictions, resulting in poor SR results by SotA (State of the Art) methods. In this paper we introduce ``Zero-Shot'' SR, which exploits the power of Deep Learning, but does not rely on prior training. We exploit the internal recurrence of information inside a single image, and train a small image-specific CNN at test time, on examples extracted solely from the input image itself. As such, it can adapt itself to different settings per image. This allows to perform SR of real old photos, noisy images, biological data, and other images where the acquisition process is unknown or non-ideal. On such images, our method outperforms SotA CNN-based SR methods, as well as previous unsupervised SR methods. To the best of our knowledge, this is the first unsupervised CNN-based SR method.

****

Detect Globally, Refine Locally: A Novel Approach to Saliency Detection

Tiantian Wang, Lihe Zhang, Shuo Wang, Huchuan Lu, Gang Yang, Xiang Ruan, Ali Borji; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3127-3135

Effective integration of contextual information is crucial for salient object detection. To achieve this, most existing methods based on 'skip' architecture mainly focus on how to integrate hierarchical features of Convolutional Neural Networks (CNNs). They simply apply concatenation or element-wise operation to incorporate high-level semantic cues and low-level detailed information. However, this can degrade the quality of predictions because cluttered and noisy information can also be passed through. To address this problem, we proposes a global Recurrent Localization Network (RLN) which exploits contextual information by the weighted response map in order to localize salient objects more accurately. % and emphasize more on useful ones. Particularly, a recurrent module is employed to progressively refine the inner structure of the CNN over multiple time steps. Moreover, to effectively recover object boundaries, we propose a local Boundary Refinement Network (BRN) to adaptively learn the local contextual information for each spatial position. The learned propagation coefficients can be used to optimally capture relations between each pixel and its neighbors. Experiments on five challenging datasets show that our approach performs favorably against all existing methods in terms of the popular evaluation metrics.

****

Beyond the Pixel-Wise Loss for Topology-Aware Delineation

Agata Mosinska, Pablo Márquez-Neila, Mateusz Kozi■ski, Pascal Fua; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3136-3145

Delineation of curvilinear structures is an important problem in Computer Vision with multiple practical applications. With the advent of Deep Learning, many current approaches on automatic delineation have focused on finding more powerful deep architectures, but have continued using the habitual pixel-wise losses such as binary cross-entropy. In this paper we claim that pixel-wise losses alone are unsuitable for this problem because of their inability to reflect the topological importance of prediction errors. Instead, we propose a new loss term that is aware of the higher-order topological features of the linear structures. We also introduce a refinement pipeline that iteratively applies the same model over the previous delineation to refine the predictions at each step while keeping the number of parameters and the complexity of the model constant. When combined with the standard pixel-wise loss, both our new loss term and iterative refinement boost the quality of the predicted delineations, in some cases almost doubling the accuracy as compared to the same classifier trained only with the binary cross-entropy. We show that our approach outperforms state-of-the-art methods on a wide range of data, from microscopy to aerial images.

****

KIPPI: KInetic Polygonal Partitioning of Images

Jean-Philippe Bauchet, Florent Lafarge; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3146-3154

Recent works showed that floating polygons can be an interesting alternative to traditional superpixels, especially for analyzing scenes with strong geometric signatures, as man-made environments. Existing algorithms produce homogeneously-sized polygons that fail to capture thin geometric structures and over-partition large uniform areas. We propose a kinetic approach that brings more flexibility on polygon shape and size. The key idea consists in progressively extending pre-detected line-segments until they meet each other. Our experiments demonstrate that output partitions both contain less polygons and better capture geometric structures than those delivered by existing methods. We also show the applicative potential of the method when used as preprocessing in object contouring.

****

Image Blind Denoising With Generative Adversarial Network Based Noise Modeling
Jingwen Chen, Jiawei Chen, Hongyang Chao, Ming Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3155-3164
In this paper, we consider a typical image blind denoising problem, which is to remove unknown noise from noisy images. As we all know, discriminative learning based methods, such as DnCNN, can achieve state-of-the-art denoising results, but they are not applicable to this problem due to the lack of paired training data. To tackle the barrier, we propose a novel two-step framework. First, a Generative Adversarial Network (GAN) is trained to estimate the noise distribution over the input noisy images and to generate noise samples. Second, the noise patches sampled from the first step are utilized to construct a paired training dataset, which is used, in turn, to train a deep Convolutional Neural Network (CNN) for denoising. Extensive experiments have been done to demonstrate the superiority of our approach in image blind denoising.

****

Multi-Scale Weighted Nuclear Norm Image Restoration
Noam Yair, Tomer Michaeli; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3165-3174
A prominent property of natural images is that groups of similar patches within them tend to lie on low-dimensional subspaces. This property has been previously used for image denoising, with particularly notable success via weighted nuclear norm minimization (WNNM). In this paper, we extend the WNNM method into a general image restoration algorithm, capable of handling arbitrary degradations (e.g. blur, missing pixels, etc.). Our approach is based on a novel regularization term which simultaneously penalizes for high weighted nuclear norm values of all the patch groups in the image. Our regularizer is isolated from the data-term, thus enabling convenient treatment of arbitrary degradations. Furthermore, it exploits the fractal property of natural images, by accounting for patch similarities also across different scales of the image. We propose a variable splitting method for solving the resulting optimization problem. This leads to an algorithm that is quite different from `plug-and-play' techniques, which solve image-restoration problems using a sequence of denoising steps. As we verify through extensive experiments, our algorithm achieves state of the art results in deblurring and inpainting, outperforming even the recent deep net based methods.

****

MoNet: Moments Embedding Network
Mengran Gou, Fei Xiong, Octavia Camps, Mario Sznaier; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3175-3183
Bilinear pooling has been recently proposed as a feature encoding layer, which can be used after the convolutional layers of a deep network, to improve performance in multiple vision tasks. Different from conventional global average pooling or fully connected layer, bilinear pooling gathers 2nd order information in a translation invariant fashion. However, a serious drawback of this family of pooling layers is their dimensionality explosion. Approximate pooling methods with compact properties have been explored towards resolving this weakness. Additionally, recent results have shown that significant performance gains can be achieved by adding 1st order information and applying matrix normalization to regularize unstable higher order information. However, combining compact pooling with matrix normalization and other order information has not been explored until n

ow. In this paper, we unify bilinear pooling and the global Gaussian embedding layers through the empirical moment matrix. In addition, we propose a novel sub-matrix square-root layer, which can be used to normalize the output of the convolution layer directly and mitigate the dimensionality problem with off-the-shelf compact pooling methods. Our experiments on three widely used fine-grained classification datasets illustrate that our proposed architecture, MoNet, can achieve similar or better performance than with the state-of-art G2DeNet. Furthermore, when combined with compact pooling technique, MoNet obtains comparable performance with encoded features with 96% less dimensions.

****

Active Fixation Control to Predict Saccade Sequences

Calden Wloka, Iuliia Kotseruba, John K. Tsotsos; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3184-3193

Visual attention is a field with a considerable history, with eye movement control and prediction forming an important subfield. Fixation modeling in the past decades has been largely dominated computationally by a number of highly influential bottom-up saliency models, such as the Itti-Koch-Niebur model. The accuracy of such models has dramatically increased recently due to deep learning. However, on static images the emphasis of these models has largely been based on non-ordered prediction of fixations through a saliency map. Very few implemented models can generate temporally ordered human-like sequences of saccades beyond an initial fixation point. Towards addressing these shortcomings we present STAR-FC, a novel multi-saccade generator based on the integration of central high-level and object-based saliency and peripheral lower-level feature-based saliency. We have evaluated our model using the CAT2000 database, successfully predicting human patterns of fixation with equivalent accuracy and quality compared to what can be achieved by using one human sequence to predict another.

****

Densely Connected Pyramid Dehazing Network

He Zhang, Vishal M. Patel; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3194-3203

We propose a new end-to-end single image dehazing method, called Densely Connected Pyramid Dehazing Network (DCPDN), which can jointly learn the transmission map, atmospheric light and dehazing all together. The end-to-end learning is achieved by directly embedding the atmospheric scattering model into the network, thereby ensuring that the proposed method strictly follows the physics-driven scattering model for dehazing. Inspired by the dense network that can maximize the information flow along features from different levels, we propose a new edge-preserving densely connected encoder-decoder structure with multi-level pyramid pooling module for estimating the transmission map. This network is optimized using a newly introduced edge-preserving loss function. To further incor- We propose a new end-to-end single image dehazing method, called Densely Connected Pyramid Dehazing Net- work (DCPDN), which can jointly learn the transmission map, atmospheric light and dehazing all together. The end- to-end learning is achieved by directly embedding the atmo- spheric scattering model into the network, thereby ensuring that the proposed method strictly follows the physics-driven scattering model for dehazing. Inspired by the dense net- work that can maximize the information flow along features from different levels, we propose a new edge-preserving densely connected encoder-decoder structure with multi- level pyramid pooling module for estimating the transmis- sion map. This network is optimized using a newly in- troduced edge-preserving loss function. To further incor- porate the mutual structural information between the esti- mated transmission map and the dehazed result, we pro- pose a joint-discriminator based on generative adversar- ial network framework to decide whether the correspond- ing dehazed image and the estimated transmission map are real or fake. An ablation study is conducted to demon- strate the effectiveness of each module evaluated at both estimated transmission map and dehazed result. Exten- sive experiments demonstrate that the proposed method achieves significant improvements over the state-of-the- art methods. Code and dataset is made available at: https://github.com/hezhangsprinter/DCPDN

****

Universal Denoising Networks : A Novel CNN Architecture for Image Denoising
Stamatios Lefkimmiatis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3204-3213

We design a novel network architecture for learning discriminative image models that are employed to efficiently tackle the problem of grayscale and color image denoising. Based on the proposed architecture, we introduce two different variants. The first network involves convolutional layers as a core component, while the second one relies instead on non-local filtering layers and thus it is able to exploit the inherent non-local self-similarity property of natural images. As opposed to most of the existing deep network approaches, which require the training of a specific model for each considered noise level, the proposed models are able to handle a wide range of noise levels using a single set of learned parameters, while they are very robust when the noise degrading the latent image does not match the statistics of the noise used during training. The latter argument is supported by results that we report on publicly available images corrupted by unknown noise and which we compare against solutions obtained by competing methods. At the same time the introduced networks achieve excellent results under additive white Gaussian noise (AWGN), which are comparable to those of the current state-of-the-art network, while they depend on a more shallow architecture with the number of trained parameters being one order of magnitude smaller. These properties make the proposed networks ideal candidates to serve as sub-solvers on restoration methods that deal with general inverse imaging problems such as deblurring, demosaicking, superresolution, etc.

****

Learning Convolutional Networks for Content-Weighted Image Compression
Mu Li, Wangmeng Zuo, Shuhang Gu, Debin Zhao, David Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3214-3223

Lossy image compression is generally formulated as a joint rate-distortion optimization problem to learn encoder, quantizer, and decoder. Due to the non-differentiable quantizer and discrete entropy estimation, it is very challenging to develop a convolutional network (CNN)-based image compression system. In this paper, motivated by that the local information content is spatially variant in an image, we suggest that: (i) the bit rate of the different parts of the image is adapted to local content, and (ii) the content-aware bit rate is allocated under the guidance of a content-weighted importance map. The sum of the importance map can thus serve as a continuous alternative of discrete entropy estimation to control compression rate. The binarizer is adopted to quantize the output of encoder and a proxy function is introduced for approximating binary operation in backward propagation to make it differentiable. The encoder, decoder, binarizer and importance map can be jointly optimized in an end-to-end manner. And a convolutional entropy encoder is further presented for lossless compression of importance map and binary codes. In low bit rate image compression, experiments show that our system significantly outperforms JPEG and JPEG 2000 by structural similarity (SSIM) index, and can produce the much better visual result with sharp edges, rich textures, and fewer artifacts.

****

Deep Video Super-Resolution Network Using Dynamic Upsampling Filters Without Explicit Motion Compensation
Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, Seon Joo Kim; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3224-3232

Video super-resolution (VSR) has become even more important recently to provide high resolution (HR) contents for ultra high definition displays. While many deep learning based VSR methods have been proposed, most of them rely heavily on the accuracy of motion estimation and compensation. We introduce a fundamentally different framework for VSR in this paper. We propose a novel end-to-end deep neural network that generates dynamic upsampling filters and a residual image, which are computed depending on the local spatio-temporal neighborhood of each pixel to avoid explicit motion compensation. With our approach, an HR image is recons tructed directly from the input image using the dynamic upsampling filters, and the fine details are added through the computed residual. Our network with the help of a new data augmentation technique can generate much sharper HR videos with temporal consistency, compared with the previous methods. We also provide analysis of our network through extensive experiments to show how the network deals with motions implicitly.

****

Erase or Fill? Deep Joint Recurrent Rain Removal and Reconstruction in Videos

Jiaying Liu, Wenhan Yang, Shuai Yang, Zongming Guo; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3233-3242

In this paper, we address the problem of video rain removal by constructing deep recurrent convolutional networks. We visit the rain removal case by considering rain occlusion regions, i.e. light transmittance of rain streaks is low. Different from additive rain streaks, in such rain occlusion regions, the details of background images are completely lost. Therefore, we propose a hybrid rain model to depict both rain streaks and occlusions. With the wealth of temporal redundancy, we build a Joint Recurrent Rain Removal and Reconstruction Network (J4R-Net) that seamlessly integrates rain degradation classification, spatial texture appearances based rain removal and temporal coherence based background details reconstruction. The rain degradation classification provides a binary map that reveals whether a location degraded by linear additive streaks or occlusions. With this side information, the gate of the recurrent unit learns to make a trade-off between rain streak removal and background details reconstruction. Extensive experiments on a series of synthetic and real videos with rain streaks verify the superiority of the proposed method over previous state-of-the-art methods.

****

Flow Guided Recurrent Neural Encoder for Video Salient Object Detection

Guanbin Li, Yuan Xie, Tianhao Wei, Keze Wang, Liang Lin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3243-3252

Image saliency detection has recently witnessed significant progress due to deep convolutional neural networks. However, extending state-of-the-art saliency detectors from image to video is challenging. The performance of salient object detection suffers from object or camera motion and the dramatic change of the appearance contrast in videos. In this paper, we present flow guided recurrent neural encoder(FGRNE), an accurate and end-to-end learning framework for video salient object detection. It works by enhancing the temporal coherence of the per-frame feature by exploiting both motion information in terms of optical flow and sequential feature evolution encoding in terms of LSTM networks. It can be considered as a universal framework to extend any FCN based static saliency detector to video salient object detection. Intensive experimental results verify the effectiveness of each part of FGRNE and confirm that our proposed method significantly outperforms state-of-the-art methods on the public benchmarks of DAVIS and FBMS.

****

Gated Fusion Network for Single Image Dehazing

Wenqi Ren, Lin Ma, Jiawei Zhang, Jinshan Pan, Xiaochun Cao, Wei Liu, Ming-Hsuan Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3253-3261

In this paper, we propose an efficient algorithm to directly restore a clear image from a hazy input. The proposed algorithm hinges on an end-to-end trainable neural network that consists of an encoder and a decoder. The encoder is exploited to capture the context of the derived input images, while the decoder is employed to estimate the contribution of each input to the final dehazed result using the learned representations attributed to the encoder. The constructed network adopts a novel fusion-based strategy which derives three inputs from an original hazy image by applying White Balance (WB), Contrast Enhancing (CE), and Gamma Correction (GC). We compute pixel-wise confidence maps based on the appearance differences between these different inputs to blend the information of the derived inputs and preserve the regions with pleasant visibility. The final dehazed image is yielded by gating the important features of the derived inputs. To train t he network, we introduce a multi-scale based approach so that the halo artifacts can be avoided. Extensive experimental results on both synthetic and real-world images demonstrate that the proposed algorithm performs favorably against the state-of-the-art algorithms.

****

Learning a Single Convolutional Super-Resolution Network for Multiple Degradations

Kai Zhang, Wangmeng Zuo, Lei Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3262-3271

Recent years have witnessed the unprecedented success of deep convolutional neural networks (CNNs) in single image super-resolution (SISR). However, existing CNN-based SISR methods mostly assume that a low-resolution (LR) image is bicubicly downsampled from a high-resolution (HR) image, thus inevitably giving rise to poor performance when the true degradation does not follow this assumption. Moreover, they lack scalability in learning a single model to non-blindly deal with multiple degradations. To address these issues, we propose a general framework with dimensionality stretching strategy that enables a single convolutional super-resolution network to take two key factors of the SISR degradation process, i.e., blur kernel and noise level, as input. Consequently, the super-resolver can handle multiple and even spatially variant degradations, which significantly improves the practicability. Extensive experimental results on synthetic and real LR images show that the proposed convolutional super-resolution network not only can produce favorable results on multiple degradations but also is computationally efficient, providing a highly effective and scalable solution to practical SISR applications.

****

Non-Blind Deblurring: Handling Kernel Uncertainty With CNNs

Subeesh Vasu, Venkatesh Reddy Maligireddy, A. N. Rajagopalan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3272-3281

Blind motion deblurring methods are primarily responsible for recovering an accurate estimate of the blur kernel. Non-blind deblurring (NBD) methods, on the other hand, attempt to faithfully restore the original image, given the blur estimate. However, NBD is quite susceptible to errors in blur kernel. In this work, we present a convolutional neural network-based approach to handle kernel uncertainty in non-blind motion deblurring. We provide multiple latent image estimates corresponding to different prior strengths obtained from a given blurry observation in order to exploit the complementarity of these inputs for improved learning. To generalize the performance to tackle arbitrary kernel noise, we train our network with a large number of real and synthetic noisy blur kernels. Our network mitigates the effects of kernel noise so as to yield detail-preserving and artifact-free restoration. Our quantitative and qualitative evaluations on benchmark datasets demonstrate that the proposed method delivers state-of-the-art results. To further underscore the benefits that can be achieved from our network, we propose two adaptations of our method to improve kernel estimates, and image deblurring quality, respectively.

****

Boundary Flow: A Siamese Network That Predicts Boundary Motion Without Training on Motion

Peng Lei, Fuxin Li, Sinisa Todorovic; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3282-3290

Using deep learning, this paper addresses the problem of joint object boundary detection and boundary motion estimation in videos, which we named boundary flow estimation. Boundary flow is an important mid-level visual cue as boundaries characterize objects' spatial extents, and the flow indicates objects' motions and interactions. Yet, most prior work on motion estimation has focused on dense object motion or feature points that may not necessarily reside on boundaries. For boundary flow estimation, we specify a new fully convolutional Siamese network (FCSN) that jointly estimates object-level boundaries in two consecutive frames. Boundary correspondences in the two frames are predicted by the same FCSN with a

new, unconventional deconvolution approach. Finally, the boundary flow estimate is improved with an edgelet-based filtering. Evaluation is conducted on three tasks: boundary detection in videos, boundary flow estimation, and optical flow estimation. On boundary detection, we achieve the state-of-the-art performance on the benchmark VSB100 dataset. On boundary flow estimation, we present the first results on the Sintel training dataset. For optical flow estimation, we run the recent approach CPM-Flow but on the augmented input with our boundary-flow matches, and achieve significant performance improvement on the Sintel benchmark.
****

Learning to See in the Dark
Chen Chen, Qifeng Chen, Jia Xu, Vladlen Koltun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3291-3300
Imaging in low light is challenging due to low photon count and low SNR. Short-exposure images suffer from noise, while long exposure can lead to blurry images and is often impractical. A variety of denoising, deblurring, and enhancement techniques have been proposed, but their effectiveness is limited in extreme conditions, such as video-rate imaging at night. To support the development of learning-based pipelines for low-light image processing, we introduce a dataset of raw short-exposure low-light images, with corresponding long-exposure reference images. Using the presented dataset, we develop a pipeline for processing low-light images, based on end-to-end training of a fully-convolutional network. The network operates directly on raw sensor data and replaces much of the traditional image processing pipeline, which tends to perform poorly on such data. We report promising results on the new dataset, analyze factors that affect performance, and highlight opportunities for future work.
****

BPGrad: Towards Global Optimality in Deep Learning via Branch and Pruning
Ziming Zhang, Yuanwei Wu, Guanghui Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3301-3309
Understanding the global optimality in deep learning (DL) has been attracting more and more attention recently. Conventional DL solvers, however, have not been developed intentionally to seek for such global optimality. In this paper we propose a novel approximation algorithm, {em BPGrad}, towards optimizing deep models globally via branch and pruning. Our BPGrad is based on the assumption of Lipschitz continuity in DL, and as a result it can adaptively determine the step size for current gradient given the history of previous updates, wherein theoretically no smaller steps can achieve the global optimality. We prove that by repeating such branch-and-pruning procedure, we can locate the global optimality within finite iterations. Empirically an efficient solver based on BPGrad for DL is proposed as well, and it outperforms conventional DL solvers such as Adagrad, Adadelta, RMSProp, and Adam in the tasks of object recognition, detection, and segmentation.
****

Perturbative Neural Networks
Felix Juefei-Xu, Vishnu Naresh Boddeti, Marios Savvides; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3310-3318
Convolutional neural networks are witnessing wide adoption in computer vision systems with numerous applications across a range of visual recognition tasks. Much of this progress is fueled through advances in convolutional neural network architectures and learning algorithms even as the basic premise of a convolutional layer has remained unchanged. In this paper, we seek to revisit the convolutional layer that has been the workhorse of state-of-the-art visual recognition models. We introduce a very simple, yet effective, module called a perturbation layer as an alternative to a convolutional layer. The perturbation layer does away with convolution in the traditional sense and instead computes its response as a weighted linear combination of non-linearly activated additive noise perturbed inputs. We demonstrate both analytically and empirically that this perturbation layer can be an effective replacement for a standard convolutional layer. Empirically, deep neural networks with perturbation layers, called Perturbative Neural

Networks (PNNs), in lieu of convolutional layers perform comparably with standard CNNs on a range of visual datasets (MNIST, CIFAR-10, PASCAL VOC, and ImageNet) with fewer parameters.

****

Unsupervised Correlation Analysis

Yedid Hoshen, Lior Wolf; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3319-3328

Linking between two data sources is a basic building block in numerous computer vision problems. In this paper, we set to answer a fundamental cognitive question: are prior correspondences necessary for linking between different domains?

One of the most popular methods for linking between domains is Canonical Correlation Analysis (CCA). All current CCA algorithms require correspondences between the views. We introduce a new method Unsupervised Correlation Analysis (UCA), which requires no prior correspondences between the two domains. The correlation maximization term in CCA is replaced by a combination of a reconstruction term (similar to autoencoders), full cycle loss, orthogonality and multiple domain confusion terms. Due to lack of supervision, the optimization leads to multiple alternative solutions with similar scores and we therefore introduce a consensus-based mechanism that is often able to recover the desired solution. Remarkably, this suffices in order to link remote domains such as text and images. We also present results on well accepted CCA benchmarks, showing that performance far exceeds other unsupervised baselines, and approaches supervised performance in some cases.

****

A Biresolution Spectral Framework for Product Quantization

Lopamudra Mukherjee, Sathya N. Ravi, Jiming Peng, Vikas Singh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3329-3338

Product quantization (PQ) (and its variants) has been effec- tively used to encode high-dimensional data into compact codes for many problems in vision. In principle, PQ decomposes the given data into a number of lower-dimensional subspaces where the quantization proceeds independently for each subspace. While the original PQ approach does not explicitly optimize for these subspaces, later proposals have argued that the performance tends to benefit significantly if such subspaces are chosen in an optimal manner. Despite such consensus, existing approaches in the literature diverge in terms of which specific properties of these subspaces are desirable and how one should proceed to solve/optimize them. Nonetheless, despite the empirical support, there is less clarity regarding the theoretical properties that underlie these experimental benefits for quantization problems in general. In this paper, we study the quantization problem in the setting where subspaces are orthogonal and show that this problem is intricately related to a specific type of spectral decomposition of the data. This insight not only opens the door to a rich body of work in spectral analysis, but also leads to distinct computational benefits. Our resultant biresolution spectral formulation captures both the subspace projection error as well as the quantization error within the same framework. After a reformulation, the core steps of our algorithm involve a simple eigen decomposition step, which can be solved efficiently. We show that our method performs very favorably against a number of state of the art methods on standard data sets.

****

Domain Adaptive Faster R-CNN for Object Detection in the Wild

Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, Luc Van Gool; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3339-3348

Object detection typically assumes that training and test data are drawn from an identical distribution, which, however, does not always hold in practice. Such a distribution mismatch will lead to a significant performance drop. In this work, we aim to improve the cross-domain robustness of object detection. We tackle the domain shift on two levels: 1) the image-level shift, such as image style, illumination, etc, and 2) the instance-level shift, such as object appearance, si ze, etc. We build our approach based on the recent state-of-the-art Faster R-CNN model, and design two domain adaptation components, on image level and instance level, to reduce the domain discrepancy. The two domain adaptation components are based on H-divergence theory, and are implemented by learning a domain classifier in adversarial training manner. The domain classifiers on different levels are further reinforced with a consistency regularization to learn a domain-invariant region proposal network (RPN) in the Faster R-CNN model. We evaluate our newly proposed approach using multiple datasets including Cityscapes, KITTI, SIM10K, etc. The results demonstrate the effectiveness of our proposed approach for robust object detection in various domain shift scenarios.

****

Low-Shot Learning With Large-Scale Diffusion

Matthijs Douze, Arthur Szlam, Bharath Hariharan, Hervé Jégou; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3349-3358

This paper considers the problem of inferring image labels from images when only a few annotated examples are available at training time. This setup is often referred to as low-shot learning, where a standard approach is to re-train the last few layers of a convolutional neural network learned on separate classes for which training examples are abundant. We consider a semi-supervised setting based on a large collection of images to support label propagation. This is possible by leveraging the recent advances on large-scale similarity graph construction. We show that despite its conceptual simplicity, scaling label propagation up to hundred millions of images leads to state of the art accuracy in the low-shot learning regime.

****

Joint Pose and Expression Modeling for Facial Expression Recognition

Feifei Zhang, Tianzhu Zhang, Qirong Mao, Changsheng Xu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3359-3368

Facial expression recognition (FER) is a challenging task due to different expressions under arbitrary poses. Most conventional approaches either perform face frontalization on a non-frontal facial image or learn separate classifiers for each pose. Different from existing methods, in this paper, we propose an end-to-end deep learning model by exploiting different poses and expressions jointly for simultaneous facial image synthesis and pose-invariant facial expression recognition. The proposed model is based on generative adversarial network (GAN) and enjoys several merits. First, the encoder-decoder structure of the generator can learn a generative and discriminative identity representation for face images. Second, the identity representation is explicitly disentangled from both expression and pose variations through the expression and pose codes. Third, our model can automatically generate face images with different expressions under arbitrary poses to enlarge and enrich the training set for FER. Quantitative and qualitative evaluations on both controlled and in-the-wild datasets demonstrate that the proposed algorithm performs favorably against state-of-the-art methods.

****

Lightweight Probabilistic Deep Networks

Jochen Gast, Stefan Roth; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3369-3378

Even though probabilistic treatments of neural networks have a long history, they have not found widespread use in practice. Sampling approaches are often too slow already for simple networks. The size of the inputs and the depth of typical CNN architectures in computer vision only compound this problem. Uncertainty in neural networks has thus been largely ignored in practice, despite the fact that it may provide important information about the reliability of predictions and the inner workings of the network. In this paper, we introduce two lightweight approaches to making supervised learning with probabilistic deep networks practical: First, we suggest probabilistic output layers for classification and regression that require only minimal changes to existing networks. Second, we employ assumed density filtering and show that activation uncertainties can be propagated in a practical fashion through the entire network, again with minor changes. Both probabilistic networks retain the predictive power of the deterministic counterpart, but yield uncertainties that correlate well with the empirical error induced by their predictions. Moreover, the robustness to adversarial examples is significantly increased.

****

Adversarially Learned One-Class Classifier for Novelty Detection

Mohammad Sabokrou, Mohammad Khalooei, Mahmood Fathy, Ehsan Adeli; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3379-3388

Novelty detection is the process of identifying the observation(s) that differ in some respect from the training observations (the target class). In reality, the novelty class is often absent during training, poorly sampled or not well defined. Therefore, one-class classifiers can efficiently model such problems. However, due to the unavailability of data from the novelty class, training an end-to-end deep network is a cumbersome task. In this paper, inspired by the success of generative adversarial networks for training deep models in unsupervised and semi-supervised settings, we propose an end-to-end architecture for one-class classification. Our architecture is composed of two deep networks, each of which trained by competing with each other while collaborating to understand the underlying concept in the target class, and then classify the testing samples. One network works as the novelty detector, while the other supports it by enhancing the inlier samples and distorting the outliers. The intuition is that the separability of the enhanced inliers and distorted outliers is much better than deciding on the original samples. The proposed framework applies to different related applications of anomaly and outlier detection in images and videos. The results on MNIST and Caltech-256 image datasets, along with the challenging UCSD Ped2 dataset for video anomaly detection illustrate that our proposed method learns the target class effectively and is superior to the baseline and state-of-the-art methods.

****

Defense Against Universal Adversarial Perturbations

Naveed Akhtar, Jian Liu, Ajmal Mian; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3389-3398

Recent advances in Deep Learning show the existence of image-agnostic quasi-imperceptible perturbations that when applied to `any' image  can fool a state-of-the-art network classifier to change its prediction about the  image label. These `Universal Adversarial Perturbations' pose a serious threat to the success of Deep Learning in practice. We present the first dedicated framework to effectively defend the networks against such perturbations. Our approach learns a Perturbation Rectifying Network (PRN) as `pre-input' layers to a targeted model, such that the targeted model needs no modification. The PRN is learned from real and synthetic image-agnostic perturbations, where an efficient method to compute the latter is also proposed. A perturbation detector is separately trained on the Discrete Cosine Transform of the input-output difference of the PRN. A query image is first passed through the PRN and verified by the detector. If a perturbation is detected, the output of the PRN is used for label prediction instead of the actual image. A rigorous evaluation shows that our framework can defend the network classifiers against unseen adversarial perturbations in the real-world scenarios with up to 96.4% success rate. The PRN also generalizes well in the sense that training for one targeted network defends another network with a comparable success rate.

****

Disentangling Factors of Variation by Mixing Them

Qiyang Hu, Attila Szabó, Tiziano Portenier, Paolo Favaro, Matthias Zwicker; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3399-3407

We propose an approach to learn image representations that consist of disentangled factors of variation without exploiting any manual labeling or data domain knowledge. A factor of variation corresponds to an image attribute that can be dis cerned consistently across a set of images, such as the pose or color of objects. Our disentangled representation consists of a concatenation of feature chunks, each chunk representing a factor of variation. It supports applications such as transferring attributes from one image to another, by simply mixing and unmixing feature chunks, and classification or retrieval based on one or several attributes, by considering a user-specified subset of feature chunks. We learn our representation without any labeling or knowledge of the data domain, using an autoencoder architecture with two novel training objectives: first, we propose an invariance objective to encourage that encoding of each attribute, and decoding of each chunk, are invariant to changes in other attributes and chunks, respectively; second, we include a classification objective, which ensures that each chunk corresponds to a consistently discernible attribute in the represented image, hence avoiding degenerate feature mappings where some chunks are completely ignored. We demonstrate the effectiveness of our approach on the MNIST, Sprites, and CelebA datasets.

****

Deformable GANs for Pose-Based Human Image Generation
Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuilière, Nicu Sebe; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3408-3416

In this paper we address the problem of generating person images   conditioned on a given pose. Specifically, given an image of a person and a target pose, we synthesize a new image of that person in the novel pose. In order to deal with pixel-to-pixel misalignments caused by the pose differences, we introduce deformable skip connections in the generator of our Generative Adversarial Network. Moreover, a nearest-neighbour loss is proposed instead of the common L1 and L2 losses in order to match the details of the generated image with the target image. We test our approach using photos of persons in different poses and we compare our method with previous work in this area showing state-of-the-art results in two benchmarks. Our method can be applied to the wider field of deformable object generation, provided that the pose of the articulated object can be extracted using a keypoint detector.

****

Hierarchical Recurrent Attention Networks for Structured Online Maps
Namdar Homayounfar, Wei-Chiu Ma, Shrinidhi Kowshika Lakshmikanth, Raquel Urtasun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3417-3426

In this paper, we tackle the problem of online road network extraction from sparse 3D point clouds. Our method is inspired by how an annotator builds a lane graph, by first identifying how many lanes there are and then drawing each one in turn. We develop a hierarchical recurrent network that attends to initial regions of a lane boundary and traces them out completely by outputting a structured polyline. We also propose a novel differentiable loss function that measures the deviation of the edges of the ground truth polylines and their predictions. This is more suitable than distances on vertices, as there exists many ways to draw equivalent polylines. We demonstrate the effectiveness of our method on a 90 km stretch of highway, and show that we can recover the right topology 92% of the time.

****

Sliced Wasserstein Distance for Learning Gaussian Mixture Models
Soheil Kolouri, Gustavo K. Rohde, Heiko Hoffmann; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3427-3436

Gaussian mixture models (GMM) are powerful parametric tools with many applications in machine learning and computer vision. Expectation maximization (EM) is the most popular algorithm for estimating the GMM parameters. However, EM guarantees only convergence to a stationary point of the log-likelihood function, which could be arbitrarily worse than the optimal solution. Inspired by the relationship between the negative log-likelihood function and the Kullback-Leibler (KL) divergence, we propose an alternative formulation for estimating the GMM parameters using the sliced Wasserstein distance, which gives rise to a new algorithm. Sp ecifically, we propose minimizing the sliced-Wasserstein distance between the mixture model and the data distribution with respect to the GMM parameters. In contrast to the KL-divergence, the energy landscape for the sliced-Wasserstein distance is more well-behaved and therefore more suitable for a stochastic gradient descent scheme to obtain the optimal GMM parameters. We show that our formulation results in parameter estimates that are more robust to random initializations and demonstrate that it can estimate high-dimensional data distributions more faithfully than the EM algorithm.

****

Aligning Infinite-Dimensional Covariance Matrices in Reproducing Kernel Hilbert Spaces for Domain Adaptation

Zhen Zhang, Mianzhi Wang, Yan Huang, Arye Nehorai; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3437-3445

Domain shift, which occurs when there is a mismatch between the distributions of training (source) and testing (target) datasets, usually results in poor performance of the trained model on the target domain. Existing algorithms typically solve this issue by reducing the distribution discrepancy in the input spaces. However, for kernel-based learning machines, performance highly depends on the statistical properties of data in reproducing kernel Hilbert spaces (RKHS). Motivated by these considerations, we propose a novel strategy for matching distributions in RKHS, which is done by aligning the RKHS covariance matrices (descriptors) across domains. This strategy is a generalization of the correlation alignment problem in Euclidean spaces to (potentially) infinite-dimensional feature spaces. In this paper, we provide two alignment approaches, for both of which we obtain closed-form expressions via kernel matrices. Furthermore, our approaches are scalable to large datasets since they can naturally handle out-of-sample instances. We conduct extensive experiments (248 domain adaptation tasks) to evaluate our approaches. Experiment results show that our approaches outperform other state-of-the-art methods in both accuracy and computationally efficiency.

****

CLEAR: Cumulative LEARning for One-Shot One-Class Image Recognition

Jedrzej Kozerawski, Matthew Turk; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3446-3455

This work addresses the novel problem of one-shot one-class classification. The goal is to estimate a classification decision boundary for a novel class based on a single image example. Our method exploits transfer learning to model the transformation from a representation of the input, extracted by a Convolutional Neural Network, to a classification decision boundary. We use a deep neural network to learn this transformation from a large labelled dataset of images and their associated class decision boundaries generated from ImageNet, and then apply the learned decision boundary to classify subsequent query images. We tested our approach on several benchmark datasets and significantly outperformed the baseline methods.

****

Local and Global Optimization Techniques in Graph-Based Clustering

Daiki Ikami, Toshihiko Yamasaki, Kiyoharu Aizawa; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3456-3464

The goal of graph-based clustering is to divide a dataset into disjoint subsets with members similar to each other from an affinity (similarity) matrix between data. The most popular method of solving graph-based clustering is spectral clustering. However, spectral clustering has drawbacks. Spectral clustering can only be applied to macro-average-based cost functions, which tend to generate undesirable small clusters. This study first introduces a novel cost function based on micro-average. We propose a local optimization method, which is widely applicable to graph-based clustering cost functions. We also propose an initial-guess-free algorithm to avoid its initialization dependency. Moreover, we present two global optimization techniques. The experimental results exhibit significant clustering performances from our proposed methods, including 100% clustering accuracy in the COIL-20 dataset.

****

Multi-Task Learning by Maximizing Statistical Dependence

Youssef A. Mejjati, Darren Cosker, Kwang In Kim; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3465-3473

We present a new multi-task learning (MTL) approach that can be applied to multiple heterogeneous task estimators. Our motivation is that the best task estimator could change depending on the task itself. For example, we may have a deep neural network for the first task and a Gaussian process for the second task. Classical MTL approaches cannot handle this case, as they require the same model or even the same parameter types for all tasks. We tackle this by considering task-specific estimators as random variables. Then, the task relationships are discovered by measuring the statistical dependence between each pair of random variables. By doing so, our model is independent of the parametric nature of each task, and is even agnostic to the existence of such parametric formulation. We compare our algorithm with existing MTL approaches on challenging real world ranking and regression datasets, and show that our approach achieves comparable or better performance without knowing the parametric form.

****

Robust Classification With Convolutional Prototype Learning

Hong-Ming Yang, Xu-Yao Zhang, Fei Yin, Cheng-Lin Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3474-3482

Convolutional neural networks (CNNs) have been widely used for image classification. Despite its high accuracies, CNN has been shown to be easily fooled by some adversarial examples, indicating that CNN is not robust enough for pattern classification. In this paper, we argue that the lack of robustness for CNN is caused by the softmax layer, which is a totally discriminative model and based on the assumption of closed world (i.e., with a fixed number of categories). To improve the robustness, we propose a novel learning framework called convolutional prototype learning (CPL). The advantage of using prototypes is that it can well handle the open world recognition problem and therefore improve the robustness. Under the framework of CPL, we design multiple classification criteria to train the network. Moreover, a prototype loss (PL) is proposed as a regularization to improve the intra-class compactness of the feature representation, which can be viewed as a generative model based on the Gaussian assumption of different classes. Experiments on several datasets demonstrate that CPL can achieve comparable or even better results than traditional CNN, and from the robustness perspective, CPL shows great advantages for both the rejection and incremental category learning tasks.

****

Generative Modeling Using the Sliced Wasserstein Distance

Ishan Deshpande, Ziyu Zhang, Alexander G. Schwing; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3483-3491

Generative Adversarial Nets (GANs) are very successful at modeling distributions from given samples, even in the high-dimensional case. However, their formulation is also known to be hard to optimize and often not stable. While this is particularly true for early GAN formulations, there has been significant empirically motivated and theoretically founded progress to improve stability, for instance, by using the Wasserstein distance rather than the Jenson-Shannon divergence. Here, we consider an alternative formulation for generative modeling based on random projections which, in its simplest form, results in a single objective rather than a saddle-point formulation. By augmenting this approach with a discriminator we improve its accuracy. We found our ap- proach to be significantly more stable compared to even the improved Wasserstein GAN. Further, unlike the traditional GAN loss, the loss formulated in our method is a good mea- sure of the actual distance between the distributions and, for the first time for GAN training, we are able to show estimates for the same.

****

Learning Time/Memory-Efficient Deep Architectures With Budgeted Super Networks

Tom Véniat, Ludovic Denoyer; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3492-3500

We propose to focus on the problem of discovering neural network architectures e fficient in terms of both prediction quality and cost. For instance, our approach is able to solve the following tasks: learn a neural network able to predict well in less than 100 milliseconds or learn an efficient model that fits in a 50 Mb memory. Our contribution is a novel family of models called Budgeted Super Networks (BSN). They are learned using gradient descent techniques applied on a budgeted learning objective function which integrates a maximum authorized cost, while making no assumption on the nature of this cost. We present a set of experiments on computer vision problems and analyze the ability of our technique to deal with three different costs: the computation cost, the memory consumption cost and a distributed computation cost. We particularly show that our model can discover neural network architectures that have a better accuracy than the ResNet and Convolutional Neural Fabrics architectures on CIFAR-10 and CIFAR-100, at a lower cost.

****

Cross-View Image Synthesis Using Conditional GANs
Krishna Regmi, Ali Borji; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3501-3510

Learning to generate natural scenes has always been a challenging task in computer vision. It is even more painstaking when the generation is conditioned on images with drastically different views. This is mainly because understanding, corresponding, and transforming appearance and semantic information across the views is not trivial. In this paper, we attempt to solve the novel problem of cross-view image synthesis, aerial to street-view and vice versa, using conditional generative adversarial networks (cGAN). Two new architectures called Crossview Fork (X-Fork) and Crossview Sequential (X-Seq) are proposed to generate scenes with resolutions of 64×64 and 256×256 pixels. X-Fork architecture has a single discriminator and a single generator. The generator hallucinates both the image and its semantic segmentation in the target view. X-Seq architecture utilizes two cGANs. The first one generates the target image which is subsequently fed to the second cGAN for generating its corresponding semantic segmentation map. The feedback from the second cGAN helps the first cGAN generate sharper images. Both of our proposed architectures learn to generate natural images as well as their semantic segmentation maps. The proposed methods show that they are able to capture and maintain the true semantics of objects in source and target views better than the traditional image-to-image translation method which considers only the visual appearance of the scene. Extensive qualitative and quantitative evaluations support the effectiveness of our frameworks, compared to two state of the art methods, for natural scene generation across drastically different views.

****

Sparse, Smart Contours to Represent and Edit Images
Tali Dekel, Chuang Gan, Dilip Krishnan, Ce Liu, William T. Freeman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3511-3520

We study the problem of reconstructing an image from information stored at contour locations. We show that high-quality reconstructions with high fidelity to the source image can be obtained from sparse input, e.g., comprising less than 6% of image pixels. This is a significant improvement over existing contour-based reconstruction methods that require much denser input to capture subtle texture information and to ensure image quality. Our model, based on generative adversarial networks, synthesizes texture and details in regions where no input information is provided. The semantic knowledge encoded into our model and the sparsity of the input allows to use contours as an intuitive interface for semantically-aware image manipulation: local edits in contour domain translate to long-range and coherent changes in pixel space. We can perform complex structural changes such as changing facial expression by simple edits of contours. Our experiments demonstrate that humans as well as a face recognition system mostly cannot distinguish between our reconstructions and the source images.

****

Anticipating Traffic Accidents With Adaptive Loss and Large-Scale Incident DB
Tomoyuki Suzuki, Hirokatsu Kataoka, Yoshimitsu Aoki, Yutaka Satoh; Proceedings o

f the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3521-3529

In this paper, we propose a novel approach for traffic accident anticipation through (i) Adaptive Loss for Early Anticipation (AdaLEA) and (ii) a large-scale self-annotated incident database. The proposed AdaLEA allows us to gradually learn an earlier anticipation as training progresses. The loss function adaptively assigns penalty weights depending on how early the model can anticipate a traffic accident at each epoch. Additionally, a new Near-miss Incident DataBase (NIDB) that contains an enormous number of traffic near-miss incidents in which the four classes of cyclist, pedestrian, vehicle, and background class are labeled is discussed. The NIDB provides joint estimations of traffic incident anticipation and risk-factor categorization. In our experimental results, we found our proposal achieved the highest scores for anticipation (99.1% mean average precision (mAP) and 4.81 sec anticipation of the average time-to-collision (ATTC), values which are +6.6% better and 2.36 sec faster than previous work) and joint estimation (62.1% (mAP) and 3.65 sec anticipation (ATTC), values which are +4.3% better and 0.70 sec faster than previous work).

****

A Minimalist Approach to Type-Agnostic Detection of Quadrics in Point Clouds

Tolga Birdal, Benjamin Busam, Nassir Navab, Slobodan Ilic, Peter Sturm; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3530-3540

This paper proposes a segmentation-free, automatic and efficient procedure to detect general geometric quadric forms in point clouds, where clutter and occlusions are inevitable. Our everyday world is dominated by man-made objects which are designed using 3D primitives (such as planes, cones, spheres, cylinders, etc.). These objects are also omnipresent in industrial environments. This gives rise to the possibility of abstracting 3D scenes through primitives, thereby positions these geometric forms as an integral part of perception and high level 3D scene understanding. As opposed to state-of-the-art, where a tailored algorithm treats each primitive type separately, we propose to encapsulate all types in a single robust detection procedure. At the center of our approach lies a closed form 3D quadric fit, operating in both primal & dual spaces and requiring as low as 4 oriented-points. Around this fit, we design a novel, local null-space voting strategy to reduce the 4-point case to 3. Voting is coupled with the famous RANSAC and makes our algorithm orders of magnitude faster than its conventional counterparts. This is the first method capable of performing a generic cross-type multi-object primitive detection in difficult scenes. Results on synthetic and real datasets support the validity of our method.

****

Facelet-Bank for Fast Portrait Manipulation

Ying-Cong Chen, Huaijia Lin, Michelle Shu, Ruiyu Li, Xin Tao, Xiaoyong Shen, Yangang Ye, Jiaya Jia; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3541-3549

Digital face manipulation has become a popular and fascinating way to touch images with the prevalence of smart phones and social networks. With a wide variety of user preferences, facial expressions, and accessories, a general and flexible model is necessary to accommodate different types of facial editing. In this paper, we propose a model to achieve this goal based on an end-to-end convolutional neural network that supports fast inference, edit-effect control, and quick partial-model update. In addition, this model learns from unpaired image sets with different attributes. Experimental results show that our framework can handle a wide range of expressions, accessories, and makeup effects. It produces high-resolution and high-quality results in fast speed.

****

Visual to Sound: Generating Natural Sound for Videos in the Wild

Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, Tamara L. Berg; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3550-3558

As two of the five traditional human senses (sight, hearing, taste, smell, and t ouch), vision and sound are basic sources through which humans understand the world. Often correlated during natural events, these two modalities combine to jointly affect human perception. In this paper, we pose the task of generating sound given visual input. Such capabilities could help enable applications in virtual reality (generating sound for virtual scenes automatically) or provide additional accessibility to images or videos for people with visual impairments. As a first step in this direction, we apply learning-based methods to generate raw waveform samples given input video frames. We evaluate our models on a dataset of videos containing a variety of sounds (such as ambient sounds and sounds from people/animals). Our experiments show that the generated sounds are fairly realistic and have good temporal synchronization with the visual inputs.

****