Holistic and Comprehensive Annotation of Clinically Significant Findings on Diverse CT Images: Learning From Radiology Reports and Label Ontology
Ke Yan, Yifan Peng, Veit Sandfort, Mohammadhadi Bagheri, Zhiyong Lu, Ronald M. Summers; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8523-8532
In radiologists' routine work, one major task is to read a medical image, e.g., a CT scan, find significant lesions, and describe them in the radiology report. In this paper, we study the lesion description or annotation problem. Given a lesion image, our aim is to predict a comprehensive set of relevant labels, such as the lesion's body part, type, and attributes, which may assist downstream fine-grained diagnosis. To address this task, we first design a deep learning module to extract relevant semantic labels from the radiology reports associated with the lesion images. With the images and text-mined labels, we propose a lesion annotation network (LesaNet) based on a multilabel convolutional neural network (CNN) to learn all labels holistically. Hierarchical relations and mutually exclusive relations between the labels are leveraged to improve the label prediction accuracy. The relations are utilized in a label expansion strategy and a reliable hard example mining algorithm. We also attach a simple score propagation layer on LesaNet to enhance recall and explore implicit relation between labels. Multi-label metric learning is combined with classification to enable interpretable prediction. We evaluated LesaNet on the public DeepLesion dataset, which contains over 32K diverse lesion images. Experiments show that LesaNet can precisely annotate the lesions using an ontology of 171 fine-grained labels with an average AUC of 0.9344.
****

Robust Histopathology Image Analysis: To Label or to Synthesize?
Le Hou, Ayush Agarwal, Dimitris Samaras, Tahsin M. Kurc, Rajarsi R. Gupta, Joel H. Saltz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8533-8542
Detection, segmentation and classification of nuclei are fundamental analysis operations in digital pathology. Existing state-of-the-art approaches demand extensive amount of supervised training data from pathologists and may still perform poorly in images from unseen tissue types. We propose an unsupervised approach for histopathology image segmentation that synthesizes heterogeneous sets of training image patches, of every tissue type. Although our synthetic patches are not always of high quality, we harness the motley crew of generated samples through a generally applicable importance sampling method. This proposed approach, for the first time, re-weighs the training loss over synthetic data so that the ideal (unbiased) generalization loss over the true data distribution is minimized. This enables us to use a random polygon generator to synthesize approximate cellular structures (i.e., nuclear masks) for which no real examples are given in many tissue types, and hence, GAN-based methods are not suited. In addition, we propose a hybrid synthesis pipeline that utilizes textures in real histopathology patches and GAN models, to tackle heterogeneity in tissue textures. Compared with existing state-of-the-art supervised models, our approach generalizes significantly better on cancer types without training data. Even in cancer types with training data, our approach achieves the same performance without supervision cost. We release code and segmentation results on over 5000 Whole Slide Images (WSI) in The Cancer Genome Atlas (TCGA) repository, a dataset that would be orders of magnitude larger than what is available today.
****

Data Augmentation Using Learned Transformations for One-Shot Medical Image Segmentation
Amy Zhao, Guha Balakrishnan, Fredo Durand, John V. Guttag, Adrian V. Dalca; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8543-8553
Image segmentation is an important task in many medical applications. Methods based on convolutional neural networks attain state-of-the-art accuracy; however, they typically rely on supervised training with large labeled datasets. Labeling medical images requires significant expertise and time, and typical hand-tuned

approaches for data augmentation fail to capture the complex variations in such images. We present an automated data augmentation method for synthesizing labeled medical images. We demonstrate our method on the task of segmenting magnetic resonance imaging (MRI) brain scans. Our method requires only a single segmented scan, and leverages other unlabeled scans in a semi-supervised approach. We learn a model of transformations from the images, and use the model along with the labeled example to synthesize additional labeled examples. Each transformation is comprised of a spatial deformation field and an intensity change, enabling the synthesis of complex effects such as variations in anatomy and image acquisition procedures. We show that training a supervised segmenter with these new examples provides significant improvements over state-of-the-art methods for one-shot biomedical image segmentation.

****

Shifting More Attention to Video Salient Object Detection

Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, Jianbing Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8554-8564

The last decade has witnessed a growing interest in video salient object detection (VSOD). However, the research community long-term lacked a well-established VSOD dataset representative of real dynamic scenes with high-quality annotations. To address this issue, we elaborately collected a visual-attention-consistent Densely Annotated VSOD (DAVSOD) dataset, which contains 226 videos with 23,938 frames that cover diverse realistic-scenes, objects, instances and motions. With corresponding real human eye-fixation data, we obtain precise ground-truths. This is the first work that explicitly emphasizes the challenge of saliency shift, i.e., the video salient object(s) may dynamically change. To further contribute the community a complete benchmark, we systematically assess 17 representative VSOD algorithms over seven existing VSOD datasets and our DAVSOD with totally 84K frames (largest-scale). Utilizing three famous metrics, we then present a comprehensive and insightful performance analysis. Furthermore, we propose a baseline model. It is equipped with a saliency shift- aware convLSTM, which can efficiently capture video saliency dynamics through learning human attention-shift behavior. Extensive experiments open up promising future directions for model development and comparison.

****

Neural Task Graphs: Generalizing to Unseen Tasks From a Single Video Demonstration

De-An Huang, Suraj Nair, Danfei Xu, Yuke Zhu, Animesh Garg, Li Fei-Fei, Silvio Savarese, Juan Carlos Niebles; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8565-8574

Our goal is to generate a policy to complete an unseen task given just a single video demonstration of the task in a given domain. We hypothesize that to successfully generalize to unseen complex tasks from a single video demonstration, it is necessary to explicitly incorporate the compositional structure of the tasks into the model. To this end, we propose Neural Task Graph (NTG) Networks, which use conjugate task graph as the intermediate representation to modularize both the video demonstration and the derived policy. We empirically show NTG achieves inter-task generalization on two complex tasks: Block Stacking in BulletPhysics and Object Collection in AI2-THOR. NTG improves data efficiency with visual input as well as achieve strong generalization without the need for dense hierarchical supervision. We further show that similar performance trends hold when applied to real-world data. We show that NTG can effectively predict task structure on the JIGSAWS surgical dataset and generalize to unseen tasks.

****

Beyond Tracking: Selecting Memory and Refining Poses for Deep Visual Odometry

Fei Xue, Xin Wang, Shunkai Li, Qiuyuan Wang, Junqiu Wang, Hongbin Zha; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8575-8583

Most previous learning-based visual odometry (VO) methods take VO as a pure tracking problem. In contrast, we present a VO framework by incorporating two additi

onal components called Memory and Refining. The Memory component preserves globa
l information by employing an adaptive and efficient selection strategy. The Ref
ining component ameliorates previous results with the contexts stored in the Mem
ory by adopting a spatial-temporal attention mechanism for feature distilling. E
xperiments on the KITTI and TUM-RGBD benchmark datasets demonstrate that our met
hod outperforms state-of-the-art learning-based methods by a large margin and pr
oduces competitive results against classic monocular VO approaches. Especially,
our model achieves outstanding performance in challenging scenarios such as text
ure-less regions and abrupt motions, where classic VO algorithms tend to fail.
****

Image Generation From Layout
Bo Zhao, Lili Meng, Weidong Yin, Leonid Sigal; Proceedings of the IEEE/CVF Co
nference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8584-8593
Despite significant recent progress on generative models, controlled generation
of images depicting multiple and complex object layouts is still a difficult pro
blem. Among the core challenges are the diversity of appearance a given object m
ay possess and, as a result, exponential set of images consistent with a specifi
ed layout. To address these challenges, we propose a novel approach for layout-b
ased image generation; we call it Layout2Im. Given the coarse spatial layout (bo
unding boxes + object categories), our model can generate a set of realistic ima
ges which have the correct objects in the desired locations. The representation
of each object is disentangled into a specified/certain part (category) and an u
nspecified/uncertain part (appearance). The category is encoded using a word emb
edding and the appearance is distilled into a low-dimensional vector sampled fro
m a normal distribution. Individual object representations are composed together
using convolutional LSTM, to obtain an encoding of the complete layout, and the
n decoded to an image. Several loss terms are introduced to encourage accurate a
nd diverse generation. The proposed Layout2Im model significantly outperforms th
e previous state of the art, boosting the best reported inception score by 24.66
% and 28.57% on the very challenging COCO-Stuff and Visual Genome datasets, resp
ectively. Extensive experiments also demonstrate our method's ability to generat
e complex and diverse images with multiple objects.
****

Multimodal Explanations by Predicting Counterfactuality in Videos
Atsushi Kanehira, Kentaro Takemoto, Sho Inayoshi, Tatsuya Harada; Proceedings
of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2
019, pp. 8594-8602
This study addresses generating counterfactual explanations with multimodal info
rmation. Our goal is not only to classify a video into a specific category, but
also to provide explanations on why it is not categorized to a specific class wi
th combinations of visual-linguistic information. Requirements that the expected
output should satisfy are referred to as counterfactuality in this paper: (1) C
ompatibility of visual-linguistic explanations, and (2) Positiveness/negativenes
s for the specific positive/negative class. Exploiting a spatio-temporal region
(tube) and an attribute as visual and linguistic explanations respectively, the
explanation model is trained to predict the counterfactuality for possible combi
nations of multimodal information in a post-hoc manner. The optimization problem
, which appears during training/inference, can be efficiently solved by insertin
g a novel neural network layer, namely the maximum subpath layer. We demonstrate
d the effectiveness of this method by comparison with a baseline of the action r
ecognition datasets extended for this task. Moreover, we provide information-the
oretical insight into the proposed method.
****

Learning to Explain With Complemental Examples
Atsushi Kanehira, Tatsuya Harada; Proceedings of the IEEE/CVF Conference on Com
puter Vision and Pattern Recognition (CVPR), 2019, pp. 8603-8611
This paper addresses the generation of explanations with visual examples. Given
an input sample, we build a system that not only classifies it to a specific cat
egory, but also outputs linguistic explanations and a set of visual examples tha
t render the decision interpretable. Focusing especially on the complementarity

of the multimodal information, i.e., linguistic and visual examples, we attempt to achieve it by maximizing the interaction information, which provides a natural definition of complementarity from an information theoretical viewpoint. We propose a novel framework to generate complemental explanations, on which the joint distribution of the variables to explain, and those to be explained is parameterized by three different neural networks: predictor, linguistic explainer, and example selector. Explanation models are trained collaboratively to maximize the interaction information to ensure the generated explanation are complemental to each other for the target. The results of experiments conducted on several data sets demonstrate the effectiveness of the proposed method.
****

HAQ: Hardware-Aware Automated Quantization With Mixed Precision
Kuan Wang, Zhijian Liu, Yujun Lin, Ji Lin, Song Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8612-8620
Model quantization is a widely used technique to compress and accelerate deep neural network (DNN) inference. Emergent DNN hardware accelerators begin to support mixed precision (1-8 bits) to further improve the computation efficiency, which raises a great challenge to find the optimal bitwidth for each layer: it requires domain experts to explore the vast design space trading off among accuracy, latency, energy, and model size, which is both time-consuming and sub-optimal. There are plenty of specialized hardware for neural networks, but little research has been done for specialized neural network optimization for a particular hardware architecture. Conventional quantization algorithm ignores the different hardware architectures and quantizes all the layers in a uniform way. In this paper, we introduce the Hardware-Aware Automated Quantization (HAQ) framework which leverages the reinforcement learning to automatically determine the quantization policy, and we take the hardware accelerator's feedback in the design loop. Rather than relying on proxy signals such as FLOPs and model size, we employ a hardware simulator to generate direct feedback signals (latency and energy) to the RL agent. Compared with conventional methods, our framework is fully automated and can specialize the quantization policy for different neural network architectures and hardware architectures. Our framework effectively reduced the latency by 1.4-1.95x and the energy consumption by 1.9x with negligible loss of accuracy compared with the fixed bitwidth (8 bits) quantization. Our framework reveals that the optimal policies on different hardware architectures (i.e., edge and cloud architectures) under different resource constraints (i.e., latency, energy and model size) are drastically different. We interpreted the implication of different quantization policies, which offer insights for both neural network architecture design and hardware architecture design.
****

Content Authentication for Neural Imaging Pipelines: End-To-End Optimization of Photo Provenance in Complex Distribution Channels
Pawel Korus, Nasir Memon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8621-8629
Forensic analysis of digital photo provenance relies on intrinsic traces left in the photograph at the time of its acquisition. Such analysis becomes unreliable after heavy post-processing, such as down-sampling and re-compression applied upon distribution in the Web. This paper explores end-to-end optimization of the entire image acquisition and distribution workflow to facilitate reliable forensic analysis at the end of the distribution channel. We demonstrate that neural imaging pipelines can be trained to replace the internals of digital cameras, and jointly optimized for high-fidelity photo development and reliable provenance analysis. In our experiments, the proposed approach increased image manipulation detection accuracy from 45% to over 90%. The findings encourage further research towards building more reliable imaging pipelines with explicit provenance-guaranteeing properties.
****

Inverse Procedural Modeling of Knitwear
Elena Trunz, Sebastian Merzbach, Jonathan Klein, Thomas Schulze, Michael Wei

nmann, Reinhard Klein; Proceedings of the IEEE/CVF Conference on Computer Visio n and Pattern Recognition (CVPR), 2019, pp. 8630-8639
The analysis and modeling of cloth has received a lot of attention in recent yea rs. While recent approaches are focused on woven cloth, we present a novel pract ical approach for the inference of more complex knitwear structures as well as t he respective knitting instructions from only a single image without attached an notations. Knitwear is produced by repeating instances of the same pattern, cons isting of grid-like arrangements of a small set of basic stitch types. Our frame work addresses the identification and localization of the occurring stitch types , which is challenging due to huge appearance variations. The resulting coarsely localized stitch types are used to infer the underlying grid structure as well as for the extraction of the knitting instruction of pattern repeats, taking int o account principles of Gestalt theory. Finally, the derived instructions allow the reproduction of the knitting structures, either as renderings or by actual k nitting, as demonstrated in several examples.
****

Estimating 3D Motion and Forces of Person-Object Interactions From Monocular Vid eo
Zongmian Li, Jiri Sedlar, Justin Carpentier, Ivan Laptev, Nicolas Mansard, Josef Sivic; Proceedings of the IEEE/CVF Conference on Computer Vision and Patte rn Recognition (CVPR), 2019, pp. 8640-8649
In this paper, we introduce a method to automatically reconstruct the 3D motion of a person interacting with an object from a single RGB video. Our method estim ates the 3D poses of the person and the object, contact positions, and forces an d torques actuated by the human limbs. The main contributions of this work are t hree-fold. First, we introduce an approach to jointly estimate the motion and th e actuation forces of the person on the manipulated object by modeling contacts and the dynamics of their interactions. This is cast as a large-scale trajectory optimization problem. Second, we develop a method to automatically recognize fr om the input video the position and timing of contacts between the person and th e object or the ground, thereby significantly simplifying the complexity of the optimization. Third, we validate our approach on a recent MoCap dataset with gr ound truth contact forces and demonstrate its performance on a new dataset of In ternet videos showing people manipulating a variety of tools in unconstrained en vironments.
****

DeepMapping: Unsupervised Map Estimation From Multiple Point Clouds
Li Ding, Chen Feng; Proceedings of the IEEE/CVF Conference on Computer Vision a nd Pattern Recognition (CVPR), 2019, pp. 8650-8659
We propose DeepMapping, a novel registration framework using deep neural network s (DNNs) as auxiliary functions to align multiple point clouds from scratch to a globally consistent frame. We use DNNs to model the highly non-convex mapping p rocess that traditionally involves hand-crafted data association, sensor pose in itialization, and global refinement. Our key novelty is that "training" these DN Ns with properly defined unsupervised losses is equivalent to solving the underl ying registration problem, but less sensitive to good initialization than ICP. O ur framework contains two DNNs: a localization network that estimates the poses for input point clouds, and a map network that models the scene structure by est imating the occupancy status of global coordinates. This allows us to convert th e registration problem to a binary occupancy classification, which can be solved efficiently using gradient-based optimization. We further show that DeepMapping can be readily extended to address the problem of Lidar SLAM by imposing geomet ric constraints between consecutive point clouds. Experiments are conducted on b oth simulated and real datasets. Qualitative and quantitative comparisons demons trate that DeepMapping often enables more robust and accurate global registratio n of multiple point clouds than existing techniques. Our code is available at ht tps://ai4ce.github.io/DeepMapping/.
****

End-To-End Interpretable Neural Motion Planner
Wenyuan Zeng, Wenjie Luo, Simon Suo, Abbas Sadat, Bin Yang, Sergio Casas,

Raquel Urtasun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8660-8669
In this paper, we propose a neural motion planner for learning to drive autonomously in complex urban scenarios that include traffic-light handling, yielding, and interactions with multiple road-users. Towards this goal, we design a holistic model that takes as input raw LIDAR data and a HD map and produces interpretable intermediate representations in the form of 3D detections and their future trajectories, as well as a cost volume defining the goodness of each position that the self-driving car can take within the planning horizon. We then sample a set of diverse physically possible trajectories and choose the one with the minimum learned cost. Importantly, our cost volume is able to naturally capture multi-modality. We demonstrate the effectiveness of our approach in real-world driving data captured in several cities in North America. Our experiments show that the learned cost volume can generate safer planning than all the baselines.
****

Divergence Triangle for Joint Training of Generator Model, Energy-Based Model, and Inferential Model
Tian Han, Erik Nijkamp, Xiaolin Fang, Mitch Hill, Song-Chun Zhu, Ying Nian Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8670-8679
This paper proposes the divergence triangle as a framework for joint training of a generator model, energy-based model and inference model. The divergence triangle is a compact and symmetric (anti-symmetric) objective function that seamlessly integrates variational learning, adversarial learning, wake-sleep algorithm, and contrastive divergence in a unified probabilistic formulation. This unification makes the processes of sampling, inference, and energy evaluation readily available without the need for costly Markov chain Monte Carlo methods. Our experiments demonstrate that the divergence triangle is capable of learning (1) an energy-based model with well-formed energy landscape, (2) direct sampling in the form of a generator network, and (3) feed-forward inference that faithfully reconstructs observed as well as synthesized data.
****

Image Deformation Meta-Networks for One-Shot Learning
Zitian Chen, Yanwei Fu, Yu-Xiong Wang, Lin Ma, Wei Liu, Martial Hebert; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8680-8689
Humans can robustly learn novel visual concepts even when images undergo various deformations and loose certain information. Mimicking the same behavior and synthesizing deformed instances of new concepts may help visual recognition systems perform better one-shot learning, i.e., learning concepts from one or few examples. Our key insight is that, while the deformed images may not be visually realistic, they still maintain critical semantic information and contribute significantly to formulating classifier decision boundaries. Inspired by the recent progress of meta-learning, we combine a meta-learner with an image deformation sub-network that produces additional training examples, and optimize both models in an end-to-end manner. The deformation sub-network learns to deform images by fusing a pair of images --- a probe image that keeps the visual content and a gallery image that diversifies the deformations. We demonstrate results on the widely used one-shot learning benchmarks (miniImageNet and ImageNet 1K Challenge datasets), which significantly outperform state-of-the-art approaches.
****

Online High Rank Matrix Completion
Jicong Fan, Madeleine Udell; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8690-8698
Recent advances in matrix completion enable data imputation in full-rank matrices by exploiting low dimensional (nonlinear) latent structure. In this paper, we develop a new model for high rank matrix completion (HRMC), together with batch and online methods to fit the model and out-of-sample extension to complete new data. The method works by (implicitly) mapping the data into a high dimensional polynomial feature space using the kernel trick; importantly, the data occupies

a low dimensional subspace in this feature space, even when the original data matrix is of full-rank. The online method can handle streaming or sequential data and adapt to non-stationary latent structure, and enjoys much lower space and time complexity than previous methods for HRMC. For example, the time complexity is reduced from O(n^3) to O(r^3), where n is the number of data points, r is the matrix rank in the feature space, and r<< n. We also provide guidance on sampling rate required for these methods to succeed. Experimental results on synthetic data and motion data validate the performance of the proposed methods.
****

Multispectral Imaging for Fine-Grained Recognition of Powders on Complex Backgrounds

Tiancheng Zhi, Bernardo R. Pires, Martial Hebert, Srinivasa G. Narasimhan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8699-8708

Hundreds of materials, such as drugs, explosives, makeup, food additives, are in the form of powder. Recognizing such powders is important for security checks, criminal identification, drug control, and quality assessment. However, powder recognition has drawn little attention in the computer vision community. Powders are hard to distinguish: they are amorphous, appear matte, have little color or texture variation and blend with surfaces they are deposited on in complex ways. To address these challenges, we present the first comprehensive dataset and approach for powder recognition using multi-spectral imaging. By using Shortwave Infrared (SWIR) multi-spectral imaging together with visible light (RGB) and Near Infrared (NIR), powders can be discriminated with reasonable accuracy. We present a method to select discriminative spectral bands to significantly reduce acquisition time while improving recognition accuracy. We propose a blending model to synthesize images of powders of various thickness deposited on a wide range of surfaces. Incorporating band selection and image synthesis, we conduct fine-grained recognition of 100 powders on complex backgrounds, and achieve 60% 70% accuracy on recognition with known powder location, and over 40% mean IoU without known location.
****

ContactDB: Analyzing and Predicting Grasp Contact via Thermal Imaging

Samarth Brahmbhatt, Cusuh Ham, Charles C. Kemp, James Hays; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8709-8719

Grasping and manipulating objects is an important human skill. Since hand-object contact is fundamental to grasping, capturing it can lead to important insights. However, observing contact through external sensors is challenging because of occlusion and the complexity of the human hand. We present ContactDB, a novel dataset of contact maps for household objects that captures the rich hand-object contact that occurs during grasping, enabled by use of a thermal camera. Participants in our study grasped 3D printed objects with a post-grasp functional intent. ContactDB includes 3750 3D meshes of 50 household objects textured with contact maps and 375K frames of synchronized RGB-D+thermal images. To the best of our knowledge, this is the first large-scale dataset that records detailed contact maps for human grasps. Analysis of this data shows the influence of functional intent and object size on grasping, the tendency to touch/avoid 'active areas', and the high frequency of palm and proximal finger contact. Finally, we train state-of-the art image translation and 3D convolution algorithms to predict diverse contact patterns from object shape. Data, code and models are available at https://contactdb.cc.gatech.edu.
****

Robust Subspace Clustering With Independent and Piecewise Identically Distributed Noise Modeling

Yuanman Li, Jiantao Zhou, Xianwei Zheng, Jinyu Tian, Yuan Yan Tang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8720-8729

Most of the existing subspace clustering (SC) frameworks assume that the noise contaminating the data is generated by an independent and identically distributed

(i.i.d.) source, where the Gaussianity is often imposed. Though these assumptio ns greatly simplify the underlying problems, they do not hold in many real-world applications. For instance, in face clustering, the noise is usually caused by random occlusions, local variations and unconstrained illuminations, which is es sentially structural and hence satisfies neither the i.i.d. property nor the Gau ssianity. In this work, we propose an independent and piecewise identically dis tributed (i.p.i.d.) noise model, where the i.i.d. property only holds locally. W e demonstrate that the i.p.i.d. model better characterizes the noise encountered in practical scenarios, and accommodates the traditional i.i.d. model as a spec ial case. Assisted by this generalized noise model, we design an information the oretic learning (ITL) framework for robust SC through a novel minimum weighted e rror entropy (MWEE) criterion. Extensive experimental results show that our prop osed SC scheme significantly outperforms the state-of-the-art competing algorith ms.

****

What Correspondences Reveal About Unknown Camera and Motion Models?

Thomas Probst, Ajad Chhatkuli, Danda Pani Paudel, Luc Van Gool; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 201 9, pp. 8730-8738

In two-view geometry, camera models and motion types are used as key knowledge a long with the image point correspondences in order to solve several key problems of 3D vision. Problems such as Structure-from-Motion (SfM) and camera self-cali bration are tackled under the assumptions of a specific camera projection model and motion type. However, these key assumptions may not be always justified, i.e .., we may often know neither the camera model nor the motion type beforehand. I n that context, one can extract only the point correspondences between images. F rom such correspondences, recovering two-view relationship --expressed by the un known camera model and motion type-- remains to be an unsolved problem. In this paper, we tackle this problem in two steps. First, we propose a method that comp utes the correct two-view relationship in the presence of noise and outliers. La ter, we study different possibilities to disambiguate the obtained relationships into camera model and motion type. By extensive experiments on both synthetic a nd real data, we verify our theory and assumptions in practical settings.

****

Self-Calibrating Deep Photometric Stereo Networks

Guanying Chen, Kai Han, Boxin Shi, Yasuyuki Matsushita, Kwan-Yee K. Wong; Pr oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8739-8747

This paper proposes an uncalibrated photometric stereo method for non-Lambertian scenes based on deep learning. Unlike previous approaches that heavily rely on assumptions of specific reflectances and light source distributions, our method is able to determine both shape and light directions of a scene with unknown arb itrary reflectances observed under unknown varying light directions. To achieve this goal, we propose a two-stage deep learning architecture, called SDPS-Net, w hich can effectively take advantage of intermediate supervision, resulting in re duced learning difficulty compared to a single-stage model. Experiments on both synthetic and real datasets show that our proposed approach significantly outper forms previous uncalibrated photometric stereo methods.

****

Argoverse: 3D Tracking and Forecasting With Rich Maps

Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak , Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, James H ays; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog nition (CVPR), 2019, pp. 8748-8757

We present Argoverse, a dataset designed to support autonomous vehicle perceptio n tasks including 3D tracking and motion forecasting. Argoverse includes sensor data collected by a fleet of autonomous vehicles in Pittsburgh and Miami as well as 3D tracking annotations, 300k extracted interesting vehicle trajectories, an d rich semantic maps. The sensor data consists of 360 degree images from 7 camer as with overlapping fields of view, forward-facing stereo imagery, 3D point clou

ds from long range LiDAR, and 6-DOF pose. Our 290km of mapped lanes contain ric
h geometric and semantic metadata which are not currently available in any publi
c dataset. All data is released under a Creative Commons license at Argoverse.or
g. In baseline experiments, we use map information such as lane direction, drive
able area, and ground height to improve the accuracy of 3D object tracking. We u
se 3D object tracking to mine for more than 300k interesting vehicle trajectorie
s to create a trajectory forecasting benchmark. Motion forecasting experiments r
anging in complexity from classical methods (k-NN) to LSTMs demonstrate that usi
ng detailed vector maps with lane-level information substantially reduces predic
tion error. Our tracking and forecasting experiments represent only a superficia
l exploration of the potential of rich maps in robotic perception. We hope that
Argoverse will enable the research community to explore these problems in greate
r depth.
****

Side Window Filtering
Hui Yin, Yuanhao Gong, Guoping Qiu; Proceedings of the IEEE/CVF Conference on
Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8758-8766
Local windows are routinely used in computer vision and almost without exception
 the center of the window is aligned with the pixels being processed. We show th
at this conventional wisdom is not universally applicable. When a pixel is on an
 edge, placing the center of the window on the pixel is one of the fundamental r
easons that cause many filtering algorithms to blur the edges. Based on this ins
ight, we propose a new Side Window Filtering (SWF) technique which aligns the wi
ndow's side or corner with the pixel being processed. The SWF technique is surpr
isingly simple yet theoretically rooted and very effective in practice. We show
that many traditional linear and nonlinear filters can be easily implemented und
er the SWF framework. Extensive analysis and experiments show that implementing
the SWF principle can significantly improve their edge preserving capabilities a
nd achieve state of the art performances in applications such as image smoothing
, denoising, enhancement, structure-preserving texture-removing, mutual-structur
e extraction, and HDR tone mapping. In addition to image filtering, we further s
how that the SWF principle can be extended to other applications involving the u
se of a local window. Using colorization by optimization as an example, we demon
strate that implementing the SWF principle can effectively prevent artifacts suc
h as color leakage associated with the conventional implementation. Given the ub
iquity of window based operations in computer vision, the new SWF technique is l
ikely to benefit many more applications.
****

Defense Against Adversarial Images Using Web-Scale Nearest-Neighbor Search
Abhimanyu Dubey, Laurens van der Maaten, Zeki Yalniz, Yixuan Li, Dhruv Mahaj
an; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn
ition (CVPR), 2019, pp. 8767-8776
A plethora of recent work has shown that convolutional networks are not robust t
o adversarial images: images that are created by perturbing a sample from the da
ta distribution as to maximize the loss on the perturbed example. In this work,
we hypothesize that adversarial perturbations move the image away from the image
 manifold in the sense that there exists no physical process that could have pro
duced the adversarial image. This hypothesis suggests that a successful defense
mechanism against adversarial images should aim to project the images back onto
the image manifold. We study such defense mechanisms, which approximate the proj
ection onto the unknown image manifold by a nearest-neighbor search against a we
b-scale image database containing tens of billions of images. Empirical evaluati
ons of this defense strategy on ImageNet suggest that it very effective in attac
k settings in which the adversary does not have access to the image database. We
 also propose two novel attack methods to break nearest-neighbor defense setting
s and show conditions under which nearest-neighbor defense fails. We perform a s
eries of ablation experiments, which suggest that there is a trade-off between r
obustness and accuracy between as we use features from deeper in the network, th
at a large index size (hundreds of millions) is crucial to get good performance,
 and that careful construction of database is crucial for robustness against nea

rest-neighbor attacks.
****

Incremental Object Learning From Contiguous Views
Stefan Stojanov, Samarth Mishra, Ngoc Anh Thai, Nikhil Dhanda, Ahmad Humayun, Chen Yu, Linda B. Smith, James M. Rehg; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8777-8786
In this work, we present CRIB (Continual Recognition Inspired by Babies), a synthetic incremental object learning environment that can produce data that models visual imagery produced by object exploration in early infancy. CRIB is coupled with a new 3D object dataset, Toys-200, that contains 200 unique toy-like object instances, and is also compatible with existing 3D datasets. Through extensive empirical evaluation of state-of-the-art incremental learning algorithms, we find the novel empirical result that repetition can significantly ameliorate the effects of catastrophic forgetting. Furthermore, we find that in certain cases repetition allows for performance approaching that of batch learning algorithms. Finally, we propose an unsupervised incremental learning task with intriguing baseline results.
****

IP102: A Large-Scale Benchmark Dataset for Insect Pest Recognition
Xiaoping Wu, Chi Zhan, Yu-Kun Lai, Ming-Ming Cheng, Jufeng Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8787-8796
Insect pests are one of the main factors affecting agricultural product yield. Accurate recognition of insect pests facilitates timely preventive measures to avoid economic losses. However, the existing datasets for the visual classification task mainly focus on common objects, e.g., flowers and dogs. This limits the application of powerful deep learning technology on specific domains like the agricultural field. In this paper, we collect a large-scale dataset named IP102 for insect pest recognition. Specifically, it contains more than 75, 000 images belonging to 102 categories, which exhibit a natural long-tailed distribution. In addition, we annotate about 19, 000 images with bounding boxes for object detection. The IP102 has a hierarchical taxonomy and the insect pests which mainly affect one specific agricultural product are grouped into the same upperlevel category. Furthermore, we perform several baseline experiments on the IP102 dataset, including handcrafted and deep feature based classification methods. Experimental results show that this dataset has the challenges of interand intra- class variance and data imbalance. We believe our IP102 will facilitate future research on practical insect pest control, fine-grained visual classification, and imbalanced learning fields. We make the dataset and pre-trained models publicly available at https://github.com/xpwu95/IP102.
****

CityFlow: A City-Scale Benchmark for Multi-Target Multi-Camera Vehicle Tracking and Re-Identification
Zheng Tang, Milind Naphade, Ming-Yu Liu, Xiaodong Yang, Stan Birchfield, Shuo Wang, Ratnesh Kumar, David Anastasiu, Jenq-Neng Hwang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8797-8806
Urban traffic optimization using traffic cameras as sensors is driving the need to advance state-of-the-art multi-target multi-camera (MTMC) tracking. This work introduces CityFlow, a city-scale traffic camera dataset consisting of more than 3 hours of synchronized HD videos from 40 cameras across 10 intersections, with the longest distance between two simultaneous cameras being 2.5 km. To the best of our knowledge, CityFlow is the largest-scale dataset in terms of spatial coverage and the number of cameras/videos in an urban environment. The dataset contains more than 200K annotated bounding boxes covering a wide range of scenes, viewing angles, vehicle models, and urban traffic flow conditions. Camera geometry and calibration information are provided to aid spatio-temporal analysis. In addition, a subset of the benchmark is made available for the task of image-based vehicle re-identification (ReID). We conducted an extensive experimental evaluation of baselines/state-of-the-art approaches in MTMC tracking, multi-target sin

gle-camera (MTSC) tracking, object detection, and image-based ReID on this datas et, analyzing the impact of different network architectures, loss functions, spa tio-temporal models and their combinations on task effectiveness. An evaluation server is launched with the release of our benchmark at the 2019 AI City Challen ge (https://www.aicitychallenge.org/) that allows researchers to compare the per formance of their newest techniques. We expect this dataset to catalyze research in this field, propel the state-of-the-art forward, and lead to deployed traffi c optimization(s) in the real world.
****

Social-IQ: A Question Answering Benchmark for Artificial Social Intelligence
Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, Louis-Philippe Morency ; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit ion (CVPR), 2019, pp. 8807-8817
As intelligent systems increasingly blend into our everyday life, artificial soc ial intelligence becomes a prominent area of research. Intelligent systems must be socially intelligent in order to comprehend human intents and maintain a rich level of interaction with humans. Human language offers a unique unconstrained approach to probe through questions and reason through answers about social situ ations. This unconstrained approach extends previous attempts to model social in telligence through numeric supervision (e.g. sentiment and emotions labels). In this paper, we introduce Social-IQ, a unconstrained benchmark specifically desig ned to train and evaluate socially intelligent technologies. By providing a rich source of open-ended questions and answers, Social-IQ opens the door to explain able social intelligence. The dataset contains rigorously annotated and validate d videos, questions and answers, as well as annotations for the complexity level of each question and answer. Social-IQ contains 1,250 natural in-the-wild socia l situations, 7,500 questions and 52,500 correct and incorrect answers. Although humans can reason about social situations with very high accuracy (95.08%), exi sting state-of-the-art computational models struggle on this task. As a result, Social-IQ brings novel challenges that will spark future research in social inte lligence modeling, visual reasoning, and multimodal question answering (QA).
****

UPSNet: A Unified Panoptic Segmentation Network
Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, R aquel Urtasun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat tern Recognition (CVPR), 2019, pp. 8818-8826
In this paper, we propose a unified panoptic segmentation network (UPSNet) for t ackling the newly proposed panoptic segmentation task. On top of a single backbo ne residual network, we first design a deformable convolution based semantic seg mentation head and a Mask R-CNN style instance segmentation head which solve the se two subtasks simultaneously. More importantly, we introduce a parameter-free panoptic head which solves the panoptic segmentation via pixel-wise classificati on. It first leverages the logits from the previous two heads and then innovativ ely expands the representation for enabling prediction of an extra unknown class which helps better resolving the conflicts between semantic and instance segmen tation. Besides, it handles the challenge caused by the varying number of instan ces and permits back propagation to the bottom modules in an end-to-end manner. Extensive experimental results on Cityscapes, COCO and our internal dataset demo nstrate that our UPSNet achieves state-of-the-art performance with much faster i nference. Code has been made available at: https://github.com/uber-research/UPSN et
****

JSIS3D: Joint Semantic-Instance Segmentation of 3D Point Clouds With Multi-Task Pointwise Networks and Multi-Value Conditional Random Fields
Quang-Hieu Pham, Thanh Nguyen, Binh-Son Hua, Gemma Roig, Sai-Kit Yeung; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2019, pp. 8827-8836
Deep learning techniques have become the to-go models for most vision-related ta sks on 2D images. However, their power has not been fully realised on several ta sks in 3D space, e.g., 3D scene understanding. In this work, we jointly address

the problems of semantic and instance segmentation of 3D point clouds. Specifica lly, we develop a multi-task pointwise network that simultaneously performs two tasks: predicting the semantic classes of 3D points and embedding the points int o high-dimensional vectors so that points of the same object instance are repres ented by similar embeddings. We then propose a multi-value conditional random fi eld model to incorporate the semantic and instance labels and formulate the prob lem of semantic and instance segmentation as jointly optimising labels in the fi eld model. The proposed method is thoroughly evaluated and compared with existin g methods on different indoor scene datasets including S3DIS and SceneNN. Experi mental results showed the robustness of the proposed joint semantic-instance seg mentation scheme over its single components. Our method also achieved state-of-t he-art performance on semantic segmentation.
****

Instance Segmentation by Jointly Optimizing Spatial Embeddings and Clustering Ba ndwidth

Davy Neven, Bert De Brabandere, Marc Proesmans, Luc Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8837-8845

Current state-of-the-art instance segmentation methods are not suited for real-t ime applications like autonomous driving, which require fast execution times at high accuracy. Although the currently dominant proposal-based methods have high accuracy, they are slow and generate masks at a fixed and low resolution. Propos al-free methods, by contrast, can generate masks at high resolution and are ofte n faster, but fail to reach the same accuracy as the proposal-based methods. In this work we propose a new clustering loss function for proposal-free instance s egmentation. The loss function pulls the spatial embeddings of pixels belonging to the same instance together and jointly learns an instance-specific clustering bandwidth, maximizing the intersection-over-union of the resulting instance mas k. When combined with a fast architecture, the network can perform instance segm entation in real-time while maintaining a high accuracy. We evaluate our method on the challenging Cityscapes benchmark and achieve top results (5% improvement over Mask R-CNN) at more than 10 fps on 2MP images.
****

DeepCO3: Deep Instance Co-Segmentation by Co-Peak Search and Co-Saliency Detecti on

Kuang-Jui Hsu, Yen-Yu Lin, Yung-Yu Chuang; Proceedings of the IEEE/CVF Confere nce on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8846-8855

In this paper, we address a new task called instance co-segmentation. Given a se t of images jointly covering object instances of a specific category, instance c o-segmentation aims to identify all of these instances and segment each of them, i.e. generating one mask for each instance. This task is important since instan ce-level segmentation is preferable for humans and many vision applications. It is also challenging because no pixel-wise annotated training data are available and the number of instances in each image is unknown. We solve this task by divi ding it into two sub-tasks, co-peak search and instance mask segmentation. In th e former sub-task, we develop a CNN-based network to detect the co-peaks as well as co-saliency maps for a pair of images. A co-peak has two endpoints, one in e ach image, that are local maxima in the response maps and similar to each other. Thereby, the two endpoints are potentially covered by a pair of instances of th e same category. In the latter subtask, we design a ranking function that takes the detected co-peaks and co-saliency maps as inputs and can select the object p roposals to produce the final results. Our method for instance co-segmentation a nd its variant for object colocalization are evaluated on four datasets, and ach ieve favorable performance against the state-of-the-art methods. The source code s and the collected datasets are available at https://github.com/KuangJuiHsu/Dee pCO3/
****

Improving Semantic Segmentation via Video Propagation and Label Relaxation

Yi Zhu, Karan Sapra, Fitsum A. Reda, Kevin J. Shih, Shawn Newsam, Andrew Ta o, Bryan Catanzaro; Proceedings of the IEEE/CVF Conference on Computer Vision a

nd Pattern Recognition (CVPR), 2019, pp. 8856-8865
Semantic segmentation requires large amounts of pixel-wise annotations to learn accurate models. In this paper, we present a video prediction-based methodology to scale up training sets by synthesizing new training samples in order to improve the accuracy of semantic segmentation networks. We exploit video prediction models' ability to predict future frames in order to also predict future labels. A joint propagation strategy is also proposed to alleviate mis-alignments in synthesized samples. We demonstrate that training segmentation models on datasets augmented by the synthesized samples leads to significant improvements in accuracy. Furthermore, we introduce a novel boundary label relaxation technique that makes training robust to annotation noise and propagation artifacts along object boundaries. Our proposed methods achieve state-of-the-art mIoUs of 83.5% on Cityscapes and 82.9% on CamVid. Our single model, without model ensembles, achieves 72.8% mIoU on the KITTI semantic segmentation test set, which surpasses the winning entry of the ROB challenge 2018.
****

Accel: A Corrective Fusion Network for Efficient Semantic Segmentation on Video
Samvit Jain, Xin Wang, Joseph E. Gonzalez; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8866-8875
We present Accel, a novel semantic video segmentation system that achieves high accuracy at low inference cost by combining the predictions of two network branches: (1) a reference branch that extracts high-detail features on a reference keyframe, and warps these features forward using frame-to-frame optical flow estimates, and (2) an update branch that computes features of adjustable quality on the current frame, performing a temporal update at each video frame. The modularity of the update branch, where feature subnetworks of varying layer depth can be inserted (e.g. ResNet-18 to ResNet-101), enables operation over a new, state-of-the-art accuracy-throughput trade-off spectrum. Over this curve, Accel models achieve both higher accuracy and faster inference times than the closest comparable single-frame segmentation networks. In general, Accel significantly outperforms previous work on efficient semantic video segmentation, correcting warping-related error that compounds on datasets with complex dynamics. Accel is end-to-end trainable and highly modular: the reference network, the optical flow network, and the update network can each be selected independently, depending on application requirements, and then jointly fine-tuned. The result is a robust, general system for fast, high-accuracy semantic segmentation on video.
****

Shape2Motion: Joint Analysis of Motion Parts and Attributes From 3D Shapes
Xiaogang Wang, Bin Zhou, Yahao Shi, Xiaowu Chen, Qinping Zhao, Kai Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8876-8884
For the task of mobility analysis of 3D shapes, we propose joint analysis for simultaneous motion part segmentation and motion attribute estimation, taking a single 3D model as input. The problem is significantly different from those tackled in the existing works which assume the availability of either a pre-existing shape segmentation or multiple 3D models in different motion states. To that end, we develop Shape2Motion which takes a single 3D point cloud as input, and jointly computes a mobility-oriented segmentation and the associated motion attributes. Shape2Motion is comprised of two deep neural networks designed for mobility proposal generation and mobility optimization, respectively. The key contribution of these networks is the novel motion-driven features and losses used in both motion part segmentation and motion attribute estimation. This is based on the observation that the movement of a functional part preserves the shape structure. We evaluate Shape2Motion with a newly proposed benchmark for mobility analysis of 3D shapes. Results demonstrate that our method achieves the state-of-the-art performance both in terms of motion part segmentation and motion attribute estimation.
****

Semantic Correlation Promoted Shape-Variant Context for Segmentation
Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, Gang Wang; Proceedings o

f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 201
9, pp. 8885-8894
Context is essential for semantic segmentation. Due to the diverse shapes of obj
ects and their complex layout in various scene images, the spatial scales and sh
apes of contexts for different objects have very large variation. It is thus ine
ffective or inefficient to aggregate various context information from a predefin
ed fixed region. In this work, we propose to generate a scale- and shape-variant
 semantic mask for each pixel to confine its contextual region. To this end, we
first propose a novel paired convolution to infer the semantic correlation of th
e pair and based on that to generate a shape mask. Using the inferred spatial sc
ope of the contextual region, we propose a shape-variant convolution, of which t
he receptive field is controlled by the shape mask that varies with the appearan
ce of input. In this way, the proposed network aggregates the context informatio
n of a pixel from its semantic-correlated region instead of a predefined fixed r
egion. Furthermore, this work also proposes a labeling denoising model to reduce
 wrong predictions caused by the noisy low-level features. Without bells and whi
stles, the proposed segmentation network achieves new state-of-the-arts consiste
ntly on the six public segmentation datasets.
****

Relation-Shape Convolutional Neural Network for Point Cloud Analysis
Yongcheng Liu, Bin Fan, Shiming Xiang, Chunhong Pan; Proceedings of the IEEE/
CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8895
-8904
Point cloud analysis is very challenging, as the shape implied in irregular poin
ts is difficult to capture. In this paper, we propose RS-CNN, namely, Relation-S
hape Convolutional Neural Network, which extends regular grid CNN to irregular c
onfiguration for point cloud analysis. The key to RS-CNN is learning from relati
on, i.e., the geometric topology constraint among points. Specifically, the conv
olutional weight for local point set is forced to learn a high-level relation ex
pression from predefined geometric priors, between a sampled point from this poi
nt set and the others. In this way, an inductive local representation with expli
cit reasoning about the spatial layout of points can be obtained, which leads to
 much shape awareness and robustness. With this convolution as a basic operator,
 RS-CNN, a hierarchical architecture can be developed to achieve contextual shap
e-aware learning for point cloud analysis. Extensive experiments on challenging
benchmarks across three tasks verify RS-CNN achieves the state of the arts.
****

Enhancing Diversity of Defocus Blur Detectors via Cross-Ensemble Network
Wenda Zhao, Bowen Zheng, Qiuhua Lin, Huchuan Lu; Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8905-891
3
Defocus blur detection (DBD) is a fundamental yet challenging topic, since the h
omogeneous region is obscure and the transition from the focused area to the unf
ocused region is gradual. Recent DBD methods make progress through exploring dee
per or wider networks with the expense of high memory and computation. In this p
aper, we propose a novel learning strategy by breaking DBD problem into multiple
 smaller defocus blur detectors and thus estimate errors can cancel out each oth
er. Our focus is the diversity enhancement via cross-ensemble network. Specifica
lly, we design an end-to-end network composed of two logical parts: feature extr
actor network (FENet) and defocus blur detector cross-ensemble network (DBD-CENe
t). FENet is constructed to extract low-level features. Then the features are fe
d into DBD-CENet containing two parallel-branches for learning two groups of def
ocus blur detectors. For each individual, we design cross-negative and self-nega
tive correlations and an error function to enhance ensemble diversity and balanc
e individual accuracy. Finally, the multiple defocus blur detectors are combined
 with a uniformly weighted average to obtain the final DBD map. Experimental res
ults indicate the superiority of our method in terms of accuracy and speed when
compared with several state-of-the-art methods.
****

BubbleNets: Learning to Select the Guidance Frame in Video Object Segmentation b

y Deep Sorting Frames
Brent A. Griffin, Jason J. Corso; Proceedings of the IEEE/CVF Conference on Com puter Vision and Pattern Recognition (CVPR), 2019, pp. 8914-8923
Semi-supervised video object segmentation has made significant progress on real and challenging videos in recent years. The current paradigm for segmentation me thods and benchmark datasets is to segment objects in video provided a single an notation in the first frame. However, we find that segmentation performance acro ss the entire video varies dramatically when selecting an alternative frame for annotation. This paper addresses the problem of learning to suggest the single b est frame across the video for user annotation-this is, in fact, never the first frame of video. We achieve this by introducing BubbleNets, a novel deep sorting network that learns to select frames using a performance-based loss function t hat enables the conversion of expansive amounts of training examples from alread y existing datasets. Using BubbleNets, we are able to achieve an 11% relative im provement in segmentation performance on the DAVIS benchmark without any changes to the underlying method of segmentation.
****

Collaborative Global-Local Networks for Memory-Efficient Segmentation of Ultra-H igh Resolution Images
Wuyang Chen, Ziyu Jiang, Zhangyang Wang, Kexin Cui, Xiaoning Qian; Proceedin gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8924-8933
Segmentation of ultra-high resolution images is increasingly demanded, yet poses significant challenges for algorithm efficiency, in particular considering the (GPU) memory limits. Current approaches either downsample an ultra-high resoluti on image, or crop it into small patches for separate processing. In either way, the loss of local fine details or global contextual information results in limit ed segmentation accuracy. We propose collaborative Global-Local Networks (GLNet) to effectively preserve both global and local information in a highly memory-ef ficient manner. GLNet is composed of a global branch and a local branch, taking the downsampled entire image and its cropped local patches as respective inputs. For segmentation, GLNet deeply fuses feature maps from two branches, capturing both the high-resolution fine structures from zoomed-in local patches and the c ontextual dependency from the downsampled input. To further resolve the potenti al class imbalance problem between background and foreground regions, we present a coarse-to-fine variant of GLNet, also being memory-efficient. Extensive exper iments and analyses have been performed on three real-world ultra-high aerial an d medical image datasets (resolution up to 30 million pixels). With only one sin gle 1080Ti GPU and less than 2GB memory used, our GLNet yields high-quality segm entation results, and achieves much more competitive accuracy-memory usage trade -offs compared to state-of-the-arts.
****

Efficient Parameter-Free Clustering Using First Neighbor Relations
Saquib Sarfraz, Vivek Sharma, Rainer Stiefelhagen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8934-89 43
We present a new clustering method in the form of a single clustering equation t hat is able to directly discover groupings in the data. The main proposition is that the first neighbor of each sample is all one needs to discover large chains and finding the groups in the data. In contrast to most existing clustering alg orithms our method does not require any hyper-parameters, distance thresholds an d/or the need to specify the number of clusters. The proposed algorithm belongs to the family of hierarchical agglomerative methods. The technique has a very lo w computational overhead, is easily scalable and applicable to large practical p roblems. Evaluation on well known datasets from different domains ranging betwee n 1077 and 8.1 million samples shows substantial performance gains when compared to the existing clustering techniques.
****

Learning Personalized Modular Network Guided by Structured Knowledge
Xiaodan Liang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat

tern Recognition (CVPR), 2019, pp. 8944-8952
The dominant deep learning approaches use a "one-size-fits-all" paradigm with th e hope that underlying characteristics of diverse inputs can be captured via a f ixed structure. They also overlook the importance of explicitly modeling feature hierarchy. However, complex real-world tasks often require discovering diverse reasoning paths for different inputs to achieve satisfying predictions, especial ly for challenging large-scale recognition tasks with complex label relations. I n this paper, we treat the structured commonsense knowledge (e.g. concept hierar chy) as the guidance of customizing more powerful and explainable network struct ures for distinct inputs, leading to dynamic and individualized inference paths. Give an off-the-shelf large network configuration, the proposed Personalized Mo dular Network (PMN) is learned by selectively activating a sequence of network m odules where each of them is designated to recognize particular levels of struct ured knowledge. Learning semantic configurations and activation of modules to al ign well with structured knowledge can be regarded as a decision-making procedur e, which is solved by a new graph-based reinforcement learning algorithm. Experi ments on three semantic segmentation tasks and classification tasks show our PMN can achieve superior performance with the reduced number of network modules whi le discovering personalized and explainable module configurations for each input .
****

A Generative Appearance Model for End-To-End Video Object Segmentation
Joakim Johnander, Martin Danelljan, Emil Brissman, Fahad Shahbaz Khan, Micha el Felsberg; Proceedings of the IEEE/CVF Conference on Computer Vision and Patte rn Recognition (CVPR), 2019, pp. 8953-8962
One of the fundamental challenges in video object segmentation is to find an eff ective representation of the target and background appearance. The best performi ng approaches resort to extensive fine-tuning of a convolutional neural network for this purpose. Besides being prohibitively expensive, this strategy cannot be truly trained end-to-end since the online fine-tuning procedure is not integrat ed into the offline training of the network. To address these issues, we propos e a network architecture that learns a powerful representation of the target and background appearance in a single forward pass. The introduced appearance modul e learns a probabilistic generative model of target and background feature distr ibutions. Given a new image, it predicts the posterior class probabilities, prov iding a highly discriminative cue, which is processed in later network modules. Both the learning and prediction stages of our appearance module are fully diffe rentiable, enabling true end-to-end training of the entire segmentation pipeline . Comprehensive experiments demonstrate the effectiveness of the proposed approa ch on three video object segmentation benchmarks. We close the gap to approaches based on online fine-tuning on DAVIS17, while operating at 15 FPS on a single G PU. Furthermore, our method outperforms all published approaches on the large-sc ale YouTube-VOS dataset.
****

A Flexible Convolutional Solver for Fast Style Transfers
Gilles Puy, Patrick Perez; Proceedings of the IEEE/CVF Conference on Computer V ision and Pattern Recognition (CVPR), 2019, pp. 8963-8972
We propose a new flexible deep convolutional neural network (convnet) to perform fast neural style transfers. Our network is trained to solve approximately, but rapidly, the artistic style transfer problem of [Gatys et al.] for arbritary st yles. While solutions already exist, our network is uniquely flexible by design: it can be manipulated at runtime to enforce new constraints on the final output . As examples, we show that it can be modified to perform tasks such as fast pho torealistic style transfer, or fast video style transfer with short term consist ency, with no retraining. This flexibility stems from the proposed architecture which is obtained by unrolling the gradient descent algorithm used in [Gatys et al.]. Regularisations added to [Gatys et al.] to solve a new task can be reporte d on-the-fly in our network, even after training.
****

Cross Domain Model Compression by Structurally Weight Sharing

Shangqian Gao, Cheng Deng, Heng Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8973-8982
Regular model compression methods focus on RGB input. While cross domain tasks demand more DNN models, each domain often needs its own model. Consequently, for such tasks, the storage cost, memory footprint and computation cost increase dramatically compared to single RGB input. Moreover, the distinct appearance and special structure in cross domain tasks make it difficult to directly apply regular compression methods on it. In this paper, thus, we propose a new robust cross domain model compression method. Specifically, the proposed method compress cross domain models by structurally weight sharing, which is achieved by regularizing the models with graph embedding at training time. Due to the channel wise weights sharing, the proposed method can reduce computation cost without specially designed algorithm. In the experiments, the proposed method achieves state of the art results on two diverse tasks: action recognition and RGB-D scene recognition.
****
TraVeLGAN: Image-To-Image Translation by Transformation Vector Learning
Matthew Amodio, Smita Krishnaswamy; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8983-8992
Interest in image-to-image translation has grown substantially in recent years with the success of unsupervised models based on the cycle-consistency assumption. The achievements of these models have been limited to a particular subset of domains where this assumption yields good results, namely homogeneous domains that are characterized by style or texture differences. We tackle the challenging problem of image-to-image translation where the domains are defined by high-level shapes and contexts, as well as including significant clutter and heterogeneity. For this purpose, we introduce a novel GAN based on preserving intra-domain vector transformations in a latent space learned by a siamese network. The traditional GAN system introduced a discriminator network to guide the generator into generating images in the target domain. To this two-network system we add a third: a siamese network that guides the generator so that each original image shares semantics with its generated version. With this new three-network system, we no longer need to constrain the generators with the ubiquitous cycle-consistency restraint or any other autoencoding regularization. As a result, the generators can learn mappings between more complex domains that differ from each other by more than just style or texture. We demonstrate our model by mapping between high-resolution, arbitrarily chosen classes from the Imagenet dataset completely without pre-processing such as cropping, centering, or filtering unrepresentative images.
****
Deep Robust Subjective Visual Property Prediction in Crowdsourcing
Qianqian Xu, Zhiyong Yang, Yangbangyan Jiang, Xiaochun Cao, Qingming Huang, Yuan Yao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8993-9001
The problem of estimating subjective visual properties (SVP) of images (e.g., Shoes A is more comfortable than B) is gaining rising attention. Due to its highly subjective nature, different annotators often exhibit different interpretations of scales when adopting absolute value tests. Therefore, recent investigations turn to collect pairwise comparisons via crowdsourcing platforms. However, crowdsourcing data usually contains outliers. For this purpose, it is desired to develop a robust model for learning SVP from crowdsourced noisy annotations. In this paper, we construct a deep SVP prediction model which not only leads to better detection of annotation outliers but also enables learning with extremely sparse annotations. Specifically, we construct a comparison multi-graph based on the collected annotations, where different labeling results correspond to edges with different directions between two vertexes. Then, we propose a generalized deep probabilistic framework which consists of an SVP prediction module and an outlier modeling module that work collaboratively and are optimized jointly. Extensive experiments on various benchmark datasets demonstrate that our new approach guarantees promising results.

****
Transferable AutoML by Model Sharing Over Grouped Datasets
Chao Xue, Junchi Yan, Rong Yan, Stephen M. Chu, Yonggang Hu, Yonghua Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9002-9011
Automated Machine Learning (AutoML) is an active area on the design of deep neural networks for specific tasks and datasets. Given the complexity of discovering new network designs, methods for speeding up the search procedure are becoming important. This paper presents a so-called transferable AutoML approach that Automated Machine Learning (AutoML) is an active area on the design of deep neural networks for specific tasks and datasets. Given the complexity of discovering new network designs, methods for speeding up the search procedure are becoming important. This paper presents a so-called transferable AutoML approach that leverages previously trained models to speed up the search process for new tasks and datasets. Our approach involves a novel meta-feature extraction technique based on the performance of benchmark models, and a dynamic dataset clustering algorithm based on Markov process and statistical hypothesis test. As such multiple models can share a common structure while with different learned parameters. The transferable AutoML can either be applied to search from scratch, search from predesigned models, or transfer from basic cells according to the difficulties of the given datasets. The experimental results on image classification show notable speedup in overall search time for multiple datasets with negligible loss in accuracy.
****
Learning Not to Learn: Training Deep Neural Networks With Biased Data
Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, Junmo Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9012-9020
We propose a novel regularization algorithm to train deep neural networks, in which data at training time is severely biased. Since a neural network efficiently learns data distribution, a network is likely to learn the bias information to categorize input data. It leads to poor performance at test time, if the bias is, in fact, irrelevant to the categorization. In this paper, we formulate a regularization loss based on mutual information between feature embedding and bias. Based on the idea of minimizing this mutual information, we propose an iterative algorithm to unlearn the bias information. We employ an additional network to predict the bias distribution and train the network adversarially against the feature embedding network. At the end of learning, the bias prediction network is not able to predict the bias not because it is poorly trained, but because the feature embedding network successfully unlearns the bias information. We also demonstrate quantitative and qualitative experimental results which show that our algorithm effectively removes the bias information from feature embedding.
****
IRLAS: Inverse Reinforcement Learning for Architecture Search
Minghao Guo, Zhao Zhong, Wei Wu, Dahua Lin, Junjie Yan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9021-9029
In this paper, we propose an inverse reinforcement learning method for architecture search (IRLAS), which trains an agent to learn to search network structures that are topologically inspired by human-designed network. Most existing architecture search approaches totally neglect the topological characteristics of architectures, which results in complicated architecture with a high inference latency. Motivated by the fact that human-designed networks are elegant in topology with a fast inference speed, we propose a mirror stimuli function inspired by biological cognition theory to extract the abstract topological knowledge of an expert human-design network (ResNeXt). To avoid raising a too strong prior over the search space, we introduce inverse reinforcement learning to train the mirror stimuli function and exploit it as a heuristic guidance for architecture search, easily generalized to different architecture search algorithms. On CIFAR-10, the best architecture searched by our proposed IRLAS achieves 2.60% error rate. For

ImageNet mobile setting, our model achieves a state-of-the-art top-1 accuracy 75 .28%, while being 2 4x faster than most auto-generated architectures. A fast ver sion of this model achieves 10% faster than MobileNetV2, while maintaining a hig her accuracy.
****

Learning for Single-Shot Confidence Calibration in Deep Neural Networks Through Stochastic Inferences
Seonguk Seo, Paul Hongsuck Seo, Bohyung Han; Proceedings of the IEEE/CVF Confe rence on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9030-9038
We propose a generic framework to calibrate accuracy and confidence of a predict ion in deep neural networks through stochastic inferences. We interpret stochast ic regularization using a Bayesian model, and analyze the relation between predi ctive uncertainty of networks and variance of the prediction scores obtained by stochastic inferences for a single example. Our empirical study shows that the a ccuracy and the score of a prediction are highly correlated with the variance of multiple stochastic inferences given by stochastic depth or dropout. Motivated by this observation, we design a novel variance-weighted confidence-integrated l oss function that is composed of two cross-entropy loss terms with respect to gr ound-truth and uniform distribution, which are balanced by variance of stochasti c prediction scores. The proposed loss function enables us to learn deep neural networks that predict confidence calibrated scores using a single inference. Our algorithm presents outstanding confidence calibration performance and improves classification accuracy when combined with two popular stochastic regularization techniques---stochastic depth and dropout---in multiple models and datasets; it alleviates overconfidence issue in deep neural networks significantly by traini ng networks to achieve prediction accuracy proportional to confidence of predict ion.
****

Attention-Based Adaptive Selection of Operations for Image Restoration in the Pr esence of Unknown Combined Distortions
Masanori Suganuma, Xing Liu, Takayuki Okatani; Proceedings of the IEEE/CVF Con ference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9039-9048
Many studies have been conducted so far on image restoration, the problem of res toring a clean image from its distorted version. There are many different types of distortion affecting image quality. Previous studies have focused on single t ypes of distortion, proposing methods for removing them. However, image quality degrades due to multiple factors in the real world. Thus, depending on applicati ons, e.g., vision for autonomous cars or surveillance cameras, we need to be abl e to deal with multiple combined distortions with unknown mixture ratios. For th is purpose, we propose a simple yet effective layer architecture of neural netwo rks. It performs multiple operations in parallel, which are weighted by an atten tion mechanism to enable selection of proper operations depending on the input. The layer can be stacked to form a deep network, which is differentiable and thu s can be trained in an end-to-end fashion by gradient descent. The experimental results show that the proposed method works better than previous methods by a go od margin on tasks of restoring images with multiple combined distortions.
****

Fully Learnable Group Convolution for Acceleration of Deep Neural Networks
Xijun Wang, Meina Kan, Shiguang Shan, Xilin Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9049-90 58
Benefitted from its great success on many tasks, deep learning is increasingly u sed on low-computational-cost devices, e.g. smartphone, embedded devices, etc. T o reduce the high computational and memory cost, in this work, we propose a full y learnable group convolution module (FLGC for short) which is quite efficient a nd can be embedded into any deep neural networks for acceleration. Specifically, our proposed method automatically learns the group structure in the training st age in a fully end-to-end manner, leading to a better structure than the existin g pre-defined, two-steps, or iterative strategies. Moreover, our method can be f urther combined with depthwise separable convolution, resulting in 5 times accel

eration than the vanilla Resnet50 on single CPU. An additional advantage is that in our FLGC the number of groups can be set as any value, but not necessarily 2^k as in most existing methods, meaning better tradeoff between accuracy and speed. As evaluated in our experiments, our method achieves better performance than existing learnable group convolution and standard group convolution when using the same number of groups.
****

EIGEN: Ecologically-Inspired GENetic Approach for Neural Network Structure Searching From Scratch
Jian Ren, Zhe Li, Jianchao Yang, Ning Xu, Tianbao Yang, David J. Foran; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9059-9068
Designing the structure of neural networks is considered one of the most challenging tasks in deep learning, especially when there is few prior knowledge about the task domain. In this paper, we propose an Ecologically-Inspired GENetic (EIGEN) approach that uses the concept of succession, extinction, mimicry, and gene duplication to search neural network structure from scratch with poorly initialized simple network and few constraints forced during the evolution, as we assume no prior knowledge about the task domain. Specifically, we first use primary succession to rapidly evolve a population of poorly initialized neural network structures into a more diverse population, followed by a secondary succession stage for fine-grained searching based on the networks from the primary succession. Extinction is applied in both stages to reduce computational cost. Mimicry is employed during the entire evolution process to help the inferior networks imitate the behavior of a superior network and gene duplication is utilized to duplicate the learned blocks of novel structures, both of which help to find better network structures. Experimental results show that our proposed approach can achieve similar or better performance compared to the existing genetic approaches with dramatically reduced computation cost. For example, the network discovered by our approach on CIFAR-100 dataset achieves 78.1% test accuracy under 120 GPU hours, compared to 77.0% test accuracy in more than 65, 536 GPU hours in [35].
****

Deep Incremental Hashing Network for Efficient Image Retrieval
Dayan Wu, Qi Dai, Jing Liu, Bo Li, Weiping Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9069-9077
Hashing has shown great potential in large-scale image retrieval due to its storage and computation efficiency, especially the recent deep supervised hashing methods. To achieve promising performance, deep supervised hashing methods require a large amount of training data from different classes. However, when images of new categories emerge, existing deep hashing methods have to retrain the CNN model and generate hash codes for all the database images again, which is impractical for large-scale retrieval system. In this paper, we propose a novel deep hashing framework, called Deep Incremental Hashing Network (DIHN), for learning hash codes in an incremental manner. DIHN learns the hash codes for the new coming images directly, while keeping the old ones unchanged. Simultaneously, a deep hash function for query set is learned by preserving the similarities between training points. Extensive experiments on two widely used image retrieval benchmarks demonstrate that the proposed DIHN framework can significantly decrease the training time while keeping the state-of-the-art retrieval accuracy.
****

Robustness via Curvature Regularization, and Vice Versa
Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Jonathan Uesato, Pascal Frossard; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9078-9086
State-of-the-art classifiers have been shown to be largely vulnerable to adversarial perturbations. One of the most effective strategies to improve robustness is adversarial training. In this paper, we investigate the effect of adversarial training on the geometry of the classification landscape and decision boundaries. We show in particular that adversarial training leads to a significant decreas

e in the curvature of the loss surface with respect to inputs, leading to a dras tically more "linear" behaviour of the network. Using a locally quadratic approx imation, we provide theoretical evidence on the existence of a strong relation b etween large robustness and small curvature. To further show the importance of r educed curvature for improving the robustness, we propose a new regularizer that directly minimizes curvature of the loss surface, and leads to adversarial robu stness that is on par with adversarial training. Besides being a more efficient and principled alternative to adversarial training, the proposed regularizer con firms our claims on the importance of exhibiting quasi-linear behavior in the vi cinity of data points in order to achieve robustness.
****

SparseFool: A Few Pixels Make a Big Difference
Apostolos Modas, Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 201 9, pp. 9087-9096
Deep Neural Networks have achieved extraordinary results on image classification tasks, but have been shown to be vulnerable to attacks with carefully crafted p erturbations of the input data. Although most attacks usually change values of m any image's pixels, it has been shown that deep networks are also vulnerable to sparse alterations of the input. However, no computationally efficient method ha s been proposed to compute sparse perturbations. In this paper, we exploit the l ow mean curvature of the decision boundary, and propose SparseFool, a geometry i nspired sparse attack that controls the sparsity of the perturbations. Extensive evaluations show that our approach computes sparse perturbations very fast, and scales efficiently to high dimensional data. We further analyze the transferabi lity and the visual effects of the perturbations, and show the existence of shar ed semantic information across the images and the networks. Finally, we show tha t adversarial training can only slightly improve the robustness against sparse a dditive perturbations computed with SparseFool.
****

Interpretable and Fine-Grained Visual Explanations for Convolutional Neural Netw orks
Jorg Wagner, Jan Mathias Kohler, Tobias Gindele, Leon Hetzel, Jakob Thaddaus Wiedemer, Sven Behnke; Proceedings of the IEEE/CVF Conference on Computer Visi on and Pattern Recognition (CVPR), 2019, pp. 9097-9107
To verify and validate networks, it is essential to gain insight into their deci sions, limitations as well as possible shortcomings of training data. In this wo rk, we propose a post-hoc, optimization based visual explanation method, which h ighlights the evidence in the input image for a specific prediction. Our approac h is based on a novel technique to defend against adversarial evidence (i.e. fau lty evidence due to artefacts) by filtering gradients during optimization. The d efense does not depend on human-tuned parameters. It enables explanations which are both fine-grained and preserve the characteristics of images, such as edges and colors. The explanations are interpretable, suited for visualizing detailed evidence and can be tested as they are valid model inputs. We qualitatively and quantitatively evaluate our approach on a multitude of models and datasets.
****

Structured Pruning of Neural Networks With Budget-Aware Regularization
Carl Lemaire, Andrew Achkar, Pierre-Marc Jodoin; Proceedings of the IEEE/CVF C onference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9108-9116
Pruning methods have shown to be effective at reducing the size of deep neural n etworks while keeping accuracy almost intact. Among the most effective methods a re those that prune a network while training it with a sparsity prior loss and l earnable dropout parameters. A shortcoming of these approaches however is that n either the size nor the inference speed of the pruned network can be controlled directly; yet this is a key feature for targeting deployment of CNNs on low-powe r hardware. To overcome this, we introduce a budgeted regularized pruning framew ork for deep CNNs. Our approach naturally fits into traditional neural network t raining as it consists of a learnable masking layer, a novel budget-aware object ive function, and the use of knowledge distillation. We also provide insights on

how to prune a residual network and how this can lead to new architectures. Experimental results reveal that CNNs pruned with our method are more accurate and less compute-hungry than state-of-the-art methods. Also, our approach is more effective at preventing accuracy collapse in case of severe pruning; this allows pruning factors of up to 16x without significant accuracy drop.
****

MBS: Macroblock Scaling for CNN Model Reduction

Yu-Hsun Lin, Chun-Nan Chou, Edward Y. Chang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9117-9125

In this paper we propose the macroblock scaling (MBS) algorithm, which can be applied to various CNN architectures to reduce their model size. MBS adaptively reduces each CNN macroblock depending on its information redundancy measured by our proposed effective flops. Empirical studies conducted with ImageNet and CIFAR-10 attest that MBS can reduce the model size of some already compact CNN models, e.g., MobileNetV2 (25.03% further reduction) and ShuffleNet (20.74%), and even ultra-deep ones such as ResNet-101 (51.67%) and ResNet-1202 (72.71%) with negligible accuracy degradation. MBS also performs better reduction at a much lower cost than the state-of-the-art optimization-based methods do. MBS's simplicity and efficiency, its flexibility to work with any CNN model, and its scalability to work with models of any depth make it an attractive choice for CNN model size reduction.
****

Fast Neural Architecture Search of Compact Semantic Segmentation Models via Auxiliary Cells

Vladimir Nekrasov, Hao Chen, Chunhua Shen, Ian Reid; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9126-9135

Automated design of neural network architectures tailored for a specific task is an extremely promising, albeit inherently difficult, avenue to explore. While most results in this domain have been achieved on image classification and language modelling problems, here we concentrate on dense per-pixel tasks, in particular, semantic image segmentation using fully convolutional networks. In contrast to the aforementioned areas, the design choices of a fully convolutional network require several changes, ranging from the sort of operations that need to be used - e.g., dilated convolutions - to a solving of a more difficult optimisation problem. In this work, we are particularly interested in searching for high-performance compact segmentation architectures, able to run in real-time using limited resources. To achieve that, we intentionally over-parameterise the architecture during the training time via a set of auxiliary cells that provide an intermediate supervisory signal and can be omitted during the evaluation phase. The design of the auxiliary cell is emitted by a controller, a neural network with the fixed structure trained using reinforcement learning. More crucially, we demonstrate how to efficiently search for these architectures within limited time and computational budgets. In particular, we rely on a progressive strategy that terminates non-promising architectures from being further trained, and on Polyak averaging coupled with knowledge distillation to speed-up the convergence. Quantitatively, in 8 GPU-days our approach discovers a set of architectures performing on-par with state-of-the-art among compact models on the semantic segmentation, pose estimation and depth prediction tasks. Code will be made available here: https://github.com/drsleep/nas-segm-pytorch
****

Generating 3D Adversarial Point Clouds

Chong Xiang, Charles R. Qi, Bo Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9136-9144

Deep neural networks are known to be vulnerable to adversarial examples which are carefully crafted instances to cause the models to make wrong predictions. While adversarial examples for 2D images and CNNs have been extensively studied, less attention has been paid to 3D data such as point clouds. Given many safety-critical 3D applications such as autonomous driving, it is important to study how adversarial point clouds could affect current deep 3D models. In this work, we p

ropose several novel algorithms to craft adversarial point clouds against PointNet, a widely used deep neural network for point cloud processing. Our algorithms work in two ways: adversarial point perturbation and adversarial point generation. For point perturbation, we shift existing points negligibly. For point generation, we generate either a set of independent and scattered points or a small number (1-3) of point clusters with meaningful shapes such as balls and airplanes which could be hidden in the human psyche. In addition, we formulate six perturbation measurement metrics tailored to the attacks in point clouds and conduct extensive experiments to evaluate the proposed algorithms on the ModelNet40 3D shape classification dataset. Overall, our attack algorithms achieve a success rate higher than 99% for all targeted attacks.

****

Partial Order Pruning: For Best Speed/Accuracy Trade-Off in Neural Architecture Search

Xin Li, Yiming Zhou, Zheng Pan, Jiashi Feng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9145-9153

Achieving good speed and accuracy trade-off on a target platform is very important in deploying deep neural networks in real world scenarios. However, most existing automatic architecture search approaches only concentrate on high performance. In this work, we propose an algorithm that can offer better speed/accuracy trade-off of searched networks, which is termed "Partial Order Pruning". It prunes the architecture search space with a partial order assumption to automatically search for the architectures with the best speed and accuracy trade-off. Our algorithm explicitly takes profile information about the inference speed on the target platform into consideration. With the proposed algorithm, we present several Dongfeng (DF) networks that provide high accuracy and fast inference speed on various application GPU platforms. By further searching decoder architectures, our DF-Seg real-time segmentation networks yield state-of-the-art speed/accuracy trade-off on both the target embedded device and the high-end GPU.

****

Memory in Memory: A Predictive Neural Network for Learning Higher-Order Non-Stationarity From Spatiotemporal Dynamics

Yunbo Wang, Jianjin Zhang, Hongyu Zhu, Mingsheng Long, Jianmin Wang, Philip S. Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9154-9162

Natural spatiotemporal processes can be highly non-stationary in many ways, e.g. the low-level non-stationarity such as spatial correlations or temporal dependencies of local pixel values; and the high-level variations such as the accumulation, deformation or dissipation of radar echoes in precipitation forecasting. From Cramer's Decomposition, any non-stationary process can be decomposed into deterministic, time-variant polynomials, plus a zero-mean stochastic term. By applying differencing operations appropriately, we may turn time-variant polynomials into a constant, making the deterministic component predictable. However, most previous recurrent neural networks for spatiotemporal prediction do not use the differential signals effectively, and their relatively simple state transition functions prevent them from learning too complicated variations in spacetime. We propose the Memory In Memory (MIM) networks and corresponding recurrent blocks for this purpose. The MIM blocks exploit the differential signals between adjacent recurrent states to model the non-stationary and approximately stationary properties in spatiotemporal dynamics with two cascaded, self-renewed memory modules. By stacking multiple MIM blocks, we could potentially handle higher-order non-stationarity. The MIM networks achieve the state-of-the-art results on four spatiotemporal prediction tasks across both synthetic and real-world datasets. We believe that the general idea of this work can be potentially applied to other time-series forecasting tasks.

****

Variational Information Distillation for Knowledge Transfer

Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D. Lawrence, Zhenwen Dai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9163-9171

Transferring knowledge from a teacher neural network pretrained on the same or a similar task to a student neural network can significantly improve the performance of the student neural network. Existing knowledge transfer approaches match the activations or the corresponding hand-crafted features of the teacher and the student networks. We propose an information-theoretic framework for knowledge transfer which formulates knowledge transfer as maximizing the mutual information between the teacher and the student networks. We compare our method with existing knowledge transfer methods on both knowledge distillation and transfer learning tasks and show that our method consistently outperforms existing methods. We further demonstrate the strength of our method on knowledge transfer across heterogeneous network architectures by transferring knowledge from a convolutional neural network (CNN) to a multi-layer perceptron (MLP) on CIFAR-10. The resulting MLP significantly outperforms the-state-of-the-art methods and it achieves similar performance to the CNN with a single convolutional layer.

****

You Look Twice: GaterNet for Dynamic Filter Selection in CNNs

Zhourong Chen, Yang Li, Samy Bengio, Si Si; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9172-9180

The concept of conditional computation for deep nets has been proposed previously to improve model performance by selectively using only parts of the model conditioned on the sample it is processing. In this paper, we investigate input-dependent dynamic filter selection in deep convolutional neural networks (CNNs). The problem is interesting because the idea of forcing different parts of the model to learn from different types of samples may help us acquire better filters in CNNs, improve the model generalization performance and potentially increase the interpretability of model behavior. We propose a novel yet simple framework called GaterNet, which involves a backbone and a gater network. The backbone network is a regular CNN that performs the major computation needed for making a prediction, while a global gater network is introduced to generate binary gates for selectively activating filters in the backbone network based on each input. Extensive experiments on CIFAR and ImageNet datasets show that our models consistently outperform the original models with a large margin. On CIFAR-10, our model also improves upon state-of-the-art results.

****

SpherePHD: Applying CNNs on a Spherical PolyHeDron Representation of 360deg Images

Yeonkun Lee, Jaeseok Jeong, Jongseob Yun, Wonjune Cho, Kuk-Jin Yoon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9181-9189

Omni-directional cameras have many advantages overconventional cameras in that they have a much wider field-of-view (FOV). Accordingly, several approaches have beenproposed recently to apply convolutional neural networks(CNNs) to omni-directional images for various visual tasks.However, most of them use image representations defined inthe Euclidean space after transforming the omni-directionalviews originally formed in the non-Euclidean space. Thistransformation leads to shape distortion due to nonuniformspatial resolving power and the loss of continuity. Theseeffects make existing convolution kernels experience diffi-culties in extracting meaningful information. This paper presents a novel method to resolve such prob-lems of applying CNNs to omni-directional images. Theproposed method utilizes a spherical polyhedron to rep-resent omni-directional views. This method minimizes thevariance of the spatial resolving power on the sphere sur-face, and includes new convolution and pooling methodsfor the proposed representation. The proposed method canalso be adopted by any existing CNN-based methods. Thefeasibility of the proposed method is demonstrated throughclassification, detection, and semantic segmentation taskswith synthetic and real datasets.

****

ESPNetv2: A Light-Weight, Power Efficient, and General Purpose Convolutional Neural Network

Sachin Mehta, Mohammad Rastegari, Linda Shapiro, Hannaneh Hajishirzi; Proceed

ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9190-9200
We introduce a light-weight, power efficient, and general purpose convolutional neural network, ESPNetv2, for modeling visual and sequential data. Our network u ses group point-wise and depth-wise dilated separable convolutions to learn repr esentations from a large effective receptive field with fewer FLOPs and paramete rs. The performance of our network is evaluated on four different tasks: (1) obj ect classification, (2) semantic segmentation, (3) object detection, and (4) lan guage modeling. Experiments on these tasks, including image classification on th e ImageNet and language modeling on the PenTree bank dataset, demonstrate the su perior performance of our method over the state-of-the-art methods. Our network outperforms ESPNet by 4-5% and has 2-4x fewer FLOPs on the PASCAL VOC and the Ci tyscapes dataset. Compared to YOLOv2 on the MS-COCO object detection, ESPNetv2 d elivers 4.4% higher accuracy with 6x fewer FLOPs. Our experiments show that ESPN etv2 is much more power efficient than existing state-of-the-art efficient metho ds including ShuffleNets and MobileNets. Our code is open-source and available a t https://github.com/sacmehta/ESPNetv2.
****

Assisted Excitation of Activations: A Learning Technique to Improve Object Detec tors
Mohammad Mahdi Derakhshani, Saeed Masoudnia, Amir Hossein Shaker, Omid Mersa, Mohammad Amin Sadeghi, Mohammad Rastegari, Babak N. Araabi; Proceedings of t he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9201-9210
We present a simple yet effective learning technique that significantly improves mAP of YOLO object detectors without compromising their speed. During network t raining, we carefully feed in localization information. We excite certain activa tions in order to help the network learn to better localize (Figure 2). In the l ater stages of training, we gradually reduce our assisted excitation to zero. We reached a new state-of-the-art in the speed-accuracy trade-off (Figure 1). Our technique improves the mAP of YOLOv2 by 3.8% and mAP of YOLOv3 by 2.2% on MSCOCO dataset. This technique is inspired from curriculum learning. It is simple and effective and it is applicable to most single-stage object detectors.
****

Exploiting Edge Features for Graph Neural Networks
Liyu Gong, Qiang Cheng; Proceedings of the IEEE/CVF Conference on Computer Visi on and Pattern Recognition (CVPR), 2019, pp. 9211-9219
Edge features contain important information about graphs. However, current state -of-the-art neural network models designed for graph learning, e.g., graph convo lutional networks (GCN) and graph attention networks (GAT), inadequately utilize edge features, especially multi-dimensional edge features. In this paper, we bu ild a new framework for a family of new graph neural network models that can mor e sufficiently exploit edge features, including those of undirected or multi-dim ensional edges. The proposed framework can consolidate current graph neural netw ork models, e.g., GCN and GAT. The proposed framework and new models have the fo llowing novelties: First, we propose to use doubly stochastic normalization of g raph edge features instead of the commonly used row or symmetric normalization a pproaches used in current graph neural networks. Second, we construct new formul as for the operations in each individual layer so that they can handle multi-dim ensional edge features. Third, for the proposed new framework, edge features are adaptive across network layers. As a result, our proposed new framework and new models are able to exploit a rich source of graph edge information. We apply ou r new models to graph node classification on several citation networks, whole gr aph classification, and regression on several molecular datasets. Compared with the current state-of-the-art methods, i.e., GCNs and GAT, our models obtain bett er performance, which testify to the importance of exploiting edge features in g raph neural networks.
****

Propagation Mechanism for Deep and Wide Neural Networks
Dejiang Xu, Mong Li Lee, Wynne Hsu; Proceedings of the IEEE/CVF Conference on

Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9220-9228
Recent deep neural networks (DNN) utilize identity mappings involving either element-wise addition or channel-wise concatenation for the propagation of these identity mappings. In this paper, we propose a new propagation mechanism called channel-wise addition (cAdd) to deal with the vanishing gradients problem without sacrificing the complexity of the learned features. Unlike channel-wise concatenation, cAdd is able to eliminate the need to store feature maps thus reducing the memory requirement. The proposed cAdd mechanism can deepen and widen existing neural architectures with fewer parameters compared to channel-wise concatenation and element-wise addition. We incorporate cAdd into state-of-the-art architectures such as ResNet, WideResNet, and CondenseNet and carry out extensive experiments on CIFAR10, CIFAR100, SVHN and ImageNet to demonstrate that cAdd-based architectures are able to achieve much higher accuracy with fewer parameters compared to their corresponding base architectures.
****

Catastrophic Child's Play: Easy to Perform, Hard to Defend Adversarial Attacks
Chih-Hui Ho, Brandon Leung, Erik Sandstrom, Yen Chang, Nuno Vasconcelos; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9229-9237
The problem of adversarial CNN attacks is considered, with an emphasis on attacks that are trivial to perform but difficult to defend. A framework for the study of such attacks is proposed, using real world object manipulations. Unlike most works in the past, this framework supports the design of attacks based on both small and large image perturbations, implemented by camera shake and pose variation. A setup is proposed for the collection of such perturbations and determination of their perceptibility. It is argued that perceptibility depends on context, and a distinction is made between imperceptible and semantically imperceptible perturbations. While the former survives image comparisons, the latter are perceptible but have no impact on human object recognition. A procedure is proposed to determine the perceptibility of perturbations using Turk experiments, and a dataset of both perturbation classes which enables replicable studies of object manipulation attacks, is assembled. Experiments using defenses based on many data sets, CNN models, and algorithms from the literature elucidate the difficulty of defending these attacks -- in fact, none of the existing defenses is found effective against them. Better results are achieved with real world data augmentation, but even this is not foolproof. These results confirm the hypothesis that current CNNs are vulnerable to attacks implementable even by a child, and that such attacks may prove difficult to defend.
****

Embedding Complementary Deep Networks for Image Classification
Qiuyu Chen, Wei Zhang, Jun Yu, Jianping Fan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9238-9247
In this paper, a deep embedding algorithm is developed to achieve higher accuracy rates on large-scale image classification. By adapting the importance of the object classes to their error rates, our deep embedding algorithm can train multiple complementary deep networks sequentially, where each of them focuses on achieving higher accuracy rates for different subsets of object classes in an easy-to-hard way. By integrating such complementary deep networks to generate an ensemble network, our deep embedding algorithm can improve the accuracy rates for the hard object classes (which initially have higher error rates) at certain degrees while effectively preserving high accuracy rates for the easy object classes. Our deep embedding algorithm has achieved higher overall accuracy rates on large scale image classification.
****

Deep Multimodal Clustering for Unsupervised Audiovisual Learning
Di Hu, Feiping Nie, Xuelong Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9248-9257
The seen birds twitter, the running cars accompany with noise, etc. These naturally audiovisual correspondences provide the possibilities to explore and understand the outside world. However, the mixed multiple objects and sounds make it in

tractable to perform efficient matching in the unconstrained environment. To set tle this problem, we propose to adequately excavate audio and visual components and perform elaborate correspondence learning among them. Concretely, a novel un supervised audiovisual learning model is proposed, named as Deep Multimodal Clus tering (DMC),that synchronously performs sets of clustering with multimodal vect ors of convolutional maps in different shared spaces for capturing multiple audi ovisual correspondences. And such integrated multimodal clustering network can b e effectively trained with max-margin loss in the end-to-end fashion. Amounts of experiments in feature evaluation and audiovisual tasks are performed. The resu lts demonstrate that DMC can learn effective unimodal representation, with which the classifier can even outperform human performance. Further, DMC shows notice able performance in sound localization, multisource detection, and audiovisual u nderstanding.

****

Dense Classification and Implanting for Few-Shot Learning

Yann Lifchitz, Yannis Avrithis, Sylvaine Picard, Andrei Bursuc; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 201 9, pp. 9258-9267

Few-shot learning for deep neural networks is a highly challenging and key probl em in many computer vision tasks. In this context, we are targeting knowledge tr ansfer from a set with abundant data to other sets with few available examples. We propose two simple and effective solutions: (i) dense classification over fea ture maps, which for the first time studies local activations in the domain of f ew-shot learning, and (ii) implanting, that is, attaching new neurons to a previ ously trained network to learn new, task-specific features. Implanting enables t raining of multiple layers in the few-shot regime, departing from most related m ethods derived from metric learning that train only the final layer. Both contri butions show consistent gains when used individually or jointly and we report st ate of the art performance on few-shot classification on miniImageNet.

****

Class-Balanced Loss Based on Effective Number of Samples

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, Serge Belongie; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 201 9, pp. 9268-9277

With the rapid increase of large-scale, real-world datasets, it becomes critical to address the problem of long-tailed data distribution (i.e., a few classes ac count for most of the data, while most classes are under-represented). Existing solutions typically adopt class re-balancing strategies such as re-sampling and re-weighting based on the number of observations for each class. In this work, w e argue that as the number of samples increases, the additional benefit of a new ly added data point will diminish. We introduce a novel theoretical framework to measure data overlap by associating with each sample a small neighboring region rather than a single point. The effective number of samples is defined as the v olume of samples and can be calculated by a simple formula (1-b^ n)/(1-b), wher e n is the number of samples and b \in [0,1) is a hyperparameter. We design a re -weighting scheme that uses the effective number of samples for each class to re -balance the loss, thereby yielding a class-balanced loss. Comprehensive experim ents are conducted on artificially induced long-tailed CIFAR datasets and large- scale datasets including ImageNet and iNaturalist. Our results show that when tr ained with the proposed class-balanced loss, the network is able to achieve sign ificant performance gains on long-tailed datasets.

****

Discovering Visual Patterns in Art Collections With Spatially-Consistent Feature Learning

Xi Shen, Alexei A. Efros, Mathieu Aubry; Proceedings of the IEEE/CVF Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9278-9287

Our goal in this paper is to discover near duplicate patterns in large collectio ns of artworks. This is harder than standard instance mining due to differences in the artistic media (oil, pastel, drawing, etc), and imperfections inherent in the copying process. Our key technical insight is to adapt a standard deep feat

ure to this task by fine-tuning it on the specific art collection using self-supervised learning. More specifically, spatial consistency between neighbouring feature matches is used as supervisory fine-tuning signal. The adapted feature leads to more accurate style invariant matching, and can be used with a standard discovery approach, based on geometric verification, to identify duplicate patterns in the dataset. The approach is evaluated on several different datasets and shows surprisingly good qualitative discovery results. For quantitative evaluation of the method, we annotated 273 near duplicate details in a dataset of 1587 artworks attributed to Jan Brueghel and his workshop. Beyond artworks, we also demonstrate improvement on localization on the Oxford5K photo dataset as well as on historical photograph localization on the Large Time Lags Location (LTLL) dataset.

****

Min-Max Statistical Alignment for Transfer Learning

Samitha Herath, Mehrtash Harandi, Basura Fernando, Richard Nock; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9288-9297

A profound idea in learning invariant features for transfer learning is to align statistical properties of the domains. In practice, this is achieved by minimizing the disparity between the domains, usually measured in terms of their statistical properties. We question the capability of this school of thought and propose to minimize the maximum disparity between domains. Furthermore, we develop an end-to-end learning scheme that enables us to benefit from the proposed min-max strategy in training deep models. We show that the min-max solution can outperform the existing statistical alignment solutions, and can compete with state-of-the-art solutions on two challenging learning tasks, namely, Unsupervised Domain Adaptation (UDA) and Zero-Shot Learning (ZSL).

****

Spatial-Aware Graph Relation Network for Large-Scale Object Detection

Hang Xu, Chenhan Jiang, Xiaodan Liang, Zhenguo Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9298-9307

How to proper encode high-order object relation in the detection system without any external knowledge? How to leverage the information between co-occurrence and locations of objects for better reasoning? These questions are key challenges towards large-scale object detection system that aims to recognize thousands of objects entangled with complex spatial and semantic relationships nowadays. Distilling key relations that may affect object recognition is crucially important since treating each region separately leads to a big performance drop when facing heavy long-tail data distributions and plenty of confusing categories. Recent works try to encode relation by constructing graphs, e.g. using handcraft linguistic knowledge between classes or implicitly learning a fully-connected graph between regions. However, the handcraft linguistic knowledge cannot be individualized for each image due to the semantic gap between linguistic and visual context while the fully-connected graph is inefficient and noisy by incorporating redundant and distracted relations/edges from irrelevant objects and backgrounds. In this work, we introduce a Spatial-aware Graph Relation Network (SGRN) to adaptive discover and incorporate key semantic and spatial relationships for reasoning over each object. Our method considers the relative location layouts and interactions among which can be easily injected into any detection pipelines to boost the performance. Specifically, our SGRN integrates a graph learner module for learning a interpatable sparse graph structure to encode relevant contextual regions and a spatial graph reasoning module with learnable spatial Gaussian kernels to perform graph inference with spatial awareness. Extensive experiments verify the effectiveness of our method, e.g. achieving around 32% improvement on VG(3000 classes) and 28% on ADE in terms of mAP.

****

Deformable ConvNets V2: More Deformable, Better Results

Xizhou Zhu, Han Hu, Stephen Lin, Jifeng Dai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9308-9316

The superior performance of Deformable Convolutional Networks arises from its ability to adapt to the geometric variations of objects. Through an examination of its adaptive behavior, we observe that while the spatial support for its neural features conforms more closely than regular ConvNets to object structure, this support may nevertheless extend well beyond the region of interest, causing features to be influenced by irrelevant image content. To address this problem, we present a reformulation of Deformable ConvNets that improves its ability to focus on pertinent image regions, through increased modeling power and stronger training. The modeling power is enhanced through a more comprehensive integration of deformable convolution within the network, and by introducing a modulation mechanism that expands the scope of deformation modeling. To effectively harness this enriched modeling capability, we guide network training via a proposed feature mimicking scheme that helps the network to learn features that reflect the object focus and classification power of R-CNN features. With the proposed contributions, this new version of Deformable ConvNets yields significant performance gains over the original model and produces leading results on the COCO benchmark for object detection and instance segmentation.
****

Interaction-And-Aggregation Network for Person Re-Identification
Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, Xilin Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9317-9326
Person re-identification (reID) benefits greatly from deep convolutional neural networks (CNNs) which learn robust feature embeddings. However, CNNs are inherently limited in modeling the large variations in person pose and scale due to their fixed geometric structures. In this paper, we propose a novel network structure, Interaction-and-Aggregation (IA), to enhance the feature representation capability of CNNs. Firstly, Spatial IA (SIA) module is introduced. It models the interdependencies between spatial features and then aggregates the correlated features corresponding to the same body parts. Unlike CNNs which extract features from fixed rectangle regions, SIA can adaptively determine the receptive fields according to the input person pose and scale. Secondly, we introduce Channel IA (CIA) module which selectively aggregates channel features to enhance the feature representation, especially for small-scale visual cues. Further, IA network can be constructed by inserting IA blocks into CNNs at any depth. We validate the effectiveness of our model for person reID by demonstrating its superiority over state-of-the-art methods on three benchmark datasets.
****

Rare Event Detection Using Disentangled Representation Learning
Ryuhei Hamaguchi, Ken Sakurada, Ryosuke Nakamura; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9327-9335
This paper presents a novel method for rare event detection from an image pair with class-imbalanced datasets. A straightforward approach for event detection tasks is to train a detection network from a large-scale dataset in an end-to-end manner. However, in many applications such as building change detection on satellite images, few positive samples are available for the training. Moreover, an image pair of scenes contains many trivial events, such as in illumination changes or background motions. These many trivial events and the class imbalance problem lead to false alarms for rare event detection. In order to overcome these difficulties, we propose a novel method to learn disentangled representations from only low-cost negative samples. The proposed method disentangles the different aspects in a pair of observations: variant and invariant factors that represent trivial events and image contents, respectively. The effectiveness of the proposed approach is verified by the quantitative evaluations on four change detection datasets, and the qualitative analysis shows that the proposed method can acquire the representations that disentangle rare events from trivial ones.
****

Shape Robust Text Detection With Progressive Scale Expansion Network
Wenhai Wang, Enze Xie, Xiang Li, Wenbo Hou, Tong Lu, Gang Yu, Shuai Shao;

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio n (CVPR), 2019, pp. 9336-9345
Scene text detection has witnessed rapid progress especially with the recent dev elopment of convolutional neural networks. However, there still exists two chall enges which prevent the algorithm into industry applications. On the one hand, m ost of the state-of-art algorithms require quadrangle bounding box which is in-a ccurate to locate the texts with arbitrary shape. On the other hand, two text in stances which are close to each other may lead to a false detection which covers both instances. Traditionally, the segmentation-based approach can relieve the first problem but usually fail to solve the second challenge. To address these t wo challenges, in this paper, we propose a novel Progressive Scale Expansion Net work (PSENet), which can precisely detect text instances with arbitrary shapes. More specifically, PSENet generates the different scale of kernels for each text instance, and gradually expands the minimal scale kernel to the text instance w ith the complete shape. Due to the fact that there are large geometrical margins among the minimal scale kernels, our method is effective to split the close tex t instances, making it easier to use segmentation-based methods to detect arbitr ary-shaped text instances. Extensive experiments on CTW1500, Total-Text, ICDAR 2 015 and ICDAR 2017 MLT validate the effectiveness of PSENet. Notably, on CTW1500 , a dataset full of long curve texts, PSENet achieves a F-measure of 74.3% at 27 FPS, and our best F-measure (82.2%) outperforms state-of-art algorithms by 6.6% . The code will be released in the future.
****

Dual Encoding for Zero-Example Video Retrieval
Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, Xun Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Reco gnition (CVPR), 2019, pp. 9346-9355
This paper attacks the challenging problem of zero-example video retrieval. In s uch a retrieval paradigm, an end user searches for unlabeled videos by ad-hoc qu eries described in natural language text with no visual example provided. Given videos as sequences of frames and queries as sequences of words, an effective se quence-to-sequence cross-modal matching is required. The majority of existing me thods are concept based, extracting relevant concepts from queries and videos an d accordingly establishing associations between the two modalities. In contrast, this paper takes a concept-free approach, proposing a dual deep encoding networ k that encodes videos and queries into powerful dense representations of their o wn. Dual encoding is conceptually simple, practically effective and end-to-end. As experiments on three benchmarks, i.e. MSR-VTT, TRECVID 2016 and 2017 Ad-hoc V ideo Search show, the proposed solution establishes a new state-of-the-art for z ero-example video retrieval.
****

MaxpoolNMS: Getting Rid of NMS Bottlenecks in Two-Stage Object Detectors
Lile Cai, Bin Zhao, Zhe Wang, Jie Lin, Chuan Sheng Foo, Mohamed Sabry Aly, Vijay Chandrasekhar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9356-9364
Modern convolutional object detectors have improved the detection accuracy signi ficantly, which in turn inspired the development of dedicated hardware accelerat ors to achieve real-time performance by exploiting inherent parallelism in the a lgorithm. Non-maximum suppression (NMS) is an indispensable operation in object detection. In stark contrast to most operations, the commonly-adopted GreedyNMS algorithm does not foster parallelism, which can be a major performance bottlene ck. In this paper, we introduce MaxpoolNMS, a parallelizable alternative to the NMS algorithm, which is based on max-pooling classification score maps. By emplo ying a novel multi-scale multi-channel max-pooling strategy, our method is 20x f aster than GreedyNMS while simultaneously achieves comparable accuracy, when qua ntified across various benchmarking datasets, i.e., MS COCO, KITTI and PASCAL VO C. Furthermore, our method is better suited for hardware-based acceleration than GreedyNMS.
****

Character Region Awareness for Text Detection

Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, Hwalsuk Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9365-9374

Scene text detection methods based on neural networks have emerged recently and have shown promising results. Previous methods trained with rigid word-level bounding boxes exhibit limitations in representing the text region in an arbitrary shape. In this paper, we propose a new scene text detection method to effectively detect text area by exploring each character and affinity between characters. To overcome the lack of individual character level annotations, our proposed framework exploits both the given character-level annotations for synthetic images and the estimated character-level ground-truths for real images acquired by the learned interim model. In order to estimate affinity between characters, the network is trained with the newly proposed representation for affinity. Extensive experiments on six benchmarks, including the TotalText and CTW-1500 datasets which contain highly curved texts in natural images, demonstrate that our character-level text detection significantly outperforms the state-of-the-art detectors. According to the results, our proposed method guarantees high flexibility in detecting complicated scene text images, such as arbitrarily-oriented, curved, or deformed texts.

****

Effective Aesthetics Prediction With Multi-Level Spatially Pooled Features

Vlad Hosu, Bastian Goldlucke, Dietmar Saupe; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9375-9383

We propose an effective deep learning approach to aesthetics quality assessment that relies on a new type of pre-trained features, and apply it to the AVA data set, the currently largest aesthetics database. While previous approaches miss some of the information in the original images, due to taking small crops, down-scaling or warping the originals during training, we propose the first method that efficiently supports full resolution images as an input, and can be trained on variable input sizes. This allows us to significantly improve upon the state of the art, increasing the Spearman rank-order correlation coefficient (SRCC) of ground-truth mean opinion scores (MOS) from the existing best reported of 0.612 to 0.756. To achieve this performance, we extract multi-level spatially pooled (MLSP) features from all convolutional blocks of a pre-trained InceptionResNet-v2 network, and train a custom shallow Convolutional Neural Network (CNN) architecture on these new features.

****

Attentive Region Embedding Network for Zero-Shot Learning

Guo-Sen Xie, Li Liu, Xiaobo Jin, Fan Zhu, Zheng Zhang, Jie Qin, Yazhou Yao, Ling Shao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9384-9393

Zero-shot learning (ZSL) aims to classify images from unseen categories, by merely utilizing seen class images as the training data. Existing works on ZSL mainly leverage the global features or learn the global regions, from which, to construct the embeddings to the semantic space. However, few of them study the discrimination power implied in local image regions (parts), which, in some sense, correspond to semantic attributes, have stronger discrimination than attributes, and can thus assist the semantic transfer between seen/unseen classes. In this paper, to discover (semantic) regions, we propose the attentive region embedding network (AREN), which is tailored to advance the ZSL task. Specifically, AREN is end-to-end trainable and consists of two network branches, i.e., the attentive region embedding (ARE) stream, and the attentive compressed second-order embedding (ACSE) stream. ARE is capable of discovering multiple part regions under the guidance of the attention and the compatibility loss. Moreover, a novel adaptive thresholding mechanism is proposed for suppressing redundant (such as background) attention regions. To further guarantee more stable semantic transfer from the perspective of second-order collaboration, ACSE is incorporated into the AREN. In the comprehensive evaluations on four benchmarks, our models achieve state-of-the-art performances under ZSL setting, and compelling results under generalized ZSL setting.

****
Explicit Spatial Encoding for Deep Local Descriptors
Arun Mukundan, Giorgos Tolias, Ondrej Chum; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9394-9403
We propose a kernelized deep local-patch descriptor based on efficient match kernels of neural network activations. Response of each receptive field is encoded together with its spatial location using explicit feature maps. Two location parametrizations, Cartesian and polar, are used to provide robustness to a different types of canonical patch misalignment. Additionally, we analyze how the conventional architecture, i.e. a fully connected layer attached after the convolutional part, encodes responses in a spatially variant way. In contrary, explicit spatial encoding is used in our descriptor, whose potential applications are not limited to local-patches. We evaluate the descriptor on standard benchmarks. Both versions, encoding 32x32 or 64x64 patches, consistently outperform all other methods on all benchmarks. The number of parameters of the model is independent of the input patch resolution.
****
Panoptic Segmentation
Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, Piotr Dollar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9404-9413
We propose and study a task we name panoptic segmentation (PS). Panoptic segmentation unifies the typically distinct tasks of semantic segmentation (assign a class label to each pixel) and instance segmentation (detect and segment each object instance). The proposed task requires generating a coherent scene segmentation that is rich and complete, an important step toward real-world vision systems. While early work in computer vision addressed related image/scene parsing tasks, these are not currently popular, possibly due to lack of appropriate metrics or associated recognition challenges. To address this, we propose a novel panoptic quality (PQ) metric that captures performance for all classes (stuff and things) in an interpretable and unified manner. Using the proposed metric, we perform a rigorous study of both human and machine performance for PS on three existing datasets, revealing interesting insights about the task. The aim of our work is to revive the interest of the community in a more unified view of image segmentation. For more analysis and up-to-date results, please check the arXiv version of the paper: \smallhttps://arxiv.org/abs/1801.00868 .
****
You Reap What You Sow: Using Videos to Generate High Precision Object Proposals for Weakly-Supervised Object Detection
Krishna Kumar Singh, Yong Jae Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9414-9422
We propose a novel way of using videos to obtain high precision object proposals for weakly-supervised object detection. Existing weakly-supervised detection approaches use off-the-shelf proposal methods like edge boxes or selective search to obtain candidate boxes. These methods provide high recall but at the expense of thousands of noisy proposals. Thus, the entire burden of finding the few relevant object regions is left to the ensuing object mining step. To mitigate this issue, we focus instead on improving the precision of the initial candidate object proposals. Since we cannot rely on localization annotations, we turn to video and leverage motion cues to automatically estimate the extent of objects to train a Weakly-supervised Region Proposal Network (W-RPN). We use the W-RPN to generate high precision object proposals, which are in turn used to re-rank high recall proposals like edge boxes or selective search according to their spatial overlap. Our W-RPN proposals lead to significant improvement in performance for state-of-the-art weakly-supervised object detection approaches on PASCAL VOC 2007 and 2012.
****
Explore-Exploit Graph Traversal for Image Retrieval
Cheng Chang, Guangwei Yu, Chundi Liu, Maksims Volkovs; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 94

23-9431
We propose a novel graph-based approach for image retrieval. Given a nearest nei ghbor graph produced by the global descriptor model, we traverse it by alternati ng between exploit and explore steps. The exploit step maximally utilizes the im mediate neighborhood of each vertex, while the explore step traverses vertices t hat are farther away in the descriptor space. By combining these two steps we ca n better capture the underlying image manifold, and successfully retrieve releva nt images that are visually dissimilar to the query. Our traversal algorithm is conceptually simple, has few tunable parameters and can be implemented with basi c data structures. This enables fast real-time inference for previously unseen q ueries with minimal memory overhead. Despite relative simplicity, we show highly competitive results on multiple public benchmarks, including the largest image retrieval dataset that is currently publicly available. Full code for this work is available here: https://github.com/layer6ai-labs/egt.
****

Dissimilarity Coefficient Based Weakly Supervised Object Detection
Aditya Arun, C.V. Jawahar, M. Pawan Kumar; Proceedings of the IEEE/CVF Confere nce on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9432-9441
We consider the problem of weakly supervised object detection, where the trainin g samples are annotated using only image-level labels that indicate the presence or absence of an object category. In order to model the uncertainty in the loca tion of the objects, we employ a dissimilarity coefficient based probabilistic l earning objective. The learning objective minimizes the difference between an an notation agnostic prediction distribution and an annotation aware conditional di stribution. The main computational challenge is the complex nature of the condit ional distribution, which consists of terms over hundreds or thousands of variab les. The complexity of the conditional distribution rules out the possibility of explicitly modeling it. Instead, we exploit the fact that deep learning framewo rks rely on stochastic optimization. This allows us to use a state of the art di screte generative model that can provide annotation consistent samples from the conditional distribution. Extensive experiments on PASCAL VOC 2007 and 2012 data sets demonstrate the efficacy of our proposed approach.
****

Kernel Transformer Networks for Compact Spherical Convolution
Yu-Chuan Su, Kristen Grauman; Proceedings of the IEEE/CVF Conference on Compute r Vision and Pattern Recognition (CVPR), 2019, pp. 9442-9451
Ideally, 360deg imagery could inherit the deep convolutional neural networks (CN Ns) already trained with great success on perspective projection images. However , existing methods to transfer CNNs from perspective to spherical images introdu ce significant computational costs and/or degradations in accuracy. We present t he Kernel Transformer Network (KTN) to efficiently transfer convolution kernels from perspective images to the equirectangular projection of 360deg images. Give n a source CNN for perspective images as input, the KTN produces a function para meterized by a polar angle and kernel as output. Given a novel 360deg image, tha t function in turn can compute convolutions for arbitrary layers and kernels as would the source CNN on the corresponding tangent plane projections. Distinct fr om all existing methods, KTNs allow model transfer: the same model can be applie d to different source CNNs with the same base architecture. This enables applica tion to multiple recognition tasks without re-training the KTN. Validating our a pproach with multiple source CNNs and datasets, we show that KTNs improve the st ate of the art for spherical convolution. KTNs successfully preserve the source CNN's accuracy, while offering transferability, scalability to typical image res olutions, and, in many cases, a substantially lower memory footprint.
****

Object Detection With Location-Aware Deformable Convolution and Backward Attenti on Filtering
Chen Zhang, Joohee Kim; Proceedings of the IEEE/CVF Conference on Computer Visi on and Pattern Recognition (CVPR), 2019, pp. 9452-9461
Multi-class and multi-scale object detection for autonomous driving is challengi ng because of the high variation in object scales and the cluttered background i

n complex street scenes. Context information and high-resolution features are th e keys to achieve a good performance in multi-scale object detection. However, c ontext information is typically unevenly distributed, and the high-resolution fe ature map also contains distractive low-level features. In this paper, we propos e a location-aware deformable convolution and a backward attention filtering to improve the detection performance. The location-aware deformable convolution ext racts the unevenly distributed context features by sampling the input from where informative context exists. Different from the original deformable convolution, the proposed method applies an individual convolutional layer on each input sam pling grid location to obtain a wide and unique receptive field for a better off set estimation. Meanwhile, the backward attention filtering module filters the h igh-resolution feature map by highlighting the informative features and suppress ing the distractive features using the semantic features from the deep layers. E xtensive experiments are conducted on the KITTI object detection and PASCAL VOC 2007 datasets. The proposed method shows an average 6% performance improvement o ver the Faster R-CNN baseline, and it has the top-3 performance on the KITTI lea derboard with the fastest processing speed.

****

Variational Prototyping-Encoder: One-Shot Learning With Prototypical Images
Junsik Kim, Tae-Hyun Oh, Seokju Lee, Fei Pan, In So Kweon; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, p p. 9462-9470

In daily life, graphic symbols, such as traffic signs and brand logos, are ubiqu itously utilized around us due to its intuitive expression beyond language bound ary. We tackle an open-set graphic symbol recognition problem by one-shot classi fication with prototypical images as a single training example for each novel cl ass. We take an approach to learn a generalizable embedding space for novel task s. We propose a new approach called variational prototyping-encoder (VPE) that l earns the image translation task from real-world input images to their correspon ding prototypical images as a meta-task. As a result, VPE learns image similarit y as well as prototypical concepts which differs from widely used metric learnin g based approaches. Our experiments with diverse datasets demonstrate that the p roposed VPE performs favorably against competing metric learning based one-shot methods. Also, our qualitative analyses show that our meta-task induces an effec tive embedding space suitable for unseen data representation.

****

Unsupervised Domain Adaptation Using Feature-Whitening and Consensus Loss
Subhankar Roy, Aliaksandr Siarohin, Enver Sangineto, Samuel Rota Bulo, Nicu Sebe, Elisa Ricci; Proceedings of the IEEE/CVF Conference on Computer Vision an d Pattern Recognition (CVPR), 2019, pp. 9471-9480

A classifier trained on a dataset seldom works on other datasets obtained under different conditions due to domain shift. This problem is commonly addressed by domain adaptation methods. In this work we introduce a novel deep learning frame work which unifies different paradigms in unsupervised domain adaptation. Specif ically, we propose domain alignment layers which implement feature whitening for the purpose of matching source and target feature distributions. Additionally, we leverage the unlabeled target data by proposing the Min-Entropy Consensus los s, which regularizes training while avoiding the adoption of many user-defined hyper-parameters. We report results on publicly available datasets, considering both digit classification and object recognition tasks. We show that, in most o f our experiments, our approach improves upon previous methods, setting new stat e-of-the-art performances.

****

FEELVOS: Fast End-To-End Embedding Learning for Video Object Segmentation
Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam, Bastian Leibe , Liang-Chieh Chen; Proceedings of the IEEE/CVF Conference on Computer Vision a nd Pattern Recognition (CVPR), 2019, pp. 9481-9490

Many of the recent successful methods for video object segmentation (VOS) are ov erly complicated, heavily rely on fine-tuning on the first frame, and/or are slo w, and are hence of limited practical use. In this work, we propose FEELVOS as a

simple and fast method which does not rely on fine-tuning. In order to segment a video, for each frame FEELVOS uses a semantic pixel-wise embedding together with a global and a local matching mechanism to transfer information from the first frame and from the previous frame of the video to the current frame. In contrast to previous work, our embedding is only used as an internal guidance of a convolutional network. Our novel dynamic segmentation head allows us to train the network, including the embedding, end-to-end for the multiple object segmentation task with a cross entropy loss. We achieve a new state of the art in video object segmentation without fine-tuning with a J&F measure of 71.5% on the DAVIS 2017 validation set. We make our code and models available at https://github.com/tensorflow/models/tree/master/research/feelvos.

****

PartNet: A Recursive Part Decomposition Network for Fine-Grained and Hierarchical Shape Segmentation

Fenggen Yu, Kun Liu, Yan Zhang, Chenyang Zhu, Kai Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9491-9500

Deep learning approaches to 3D shape segmentation are typically formulated as a multi-class labeling problem. These models are trained for a fixed set of labels, which greatly limits their flexibility and adaptivity. We opt for top-down recursive decomposition and develop the first deep learning model for hierarchical segmentation of 3D shapes, based on recursive neural networks. Starting from a full shape represented as a point cloud, our model performs recursive binary decomposition, where the decomposition network at all nodes in the hierarchy share weights. At each node, a node classifier is trained to determine the type (adjacency or symmetry) and stopping criteria of its decomposition. The features extracted in higher level nodes are recursively propagated to lower level ones. Thus, the meaningful decompositions in higher levels provide strong contextual cues constraining the segmentations in lower levels. Meanwhile, to increase the segmentation accuracy at each node, we enhance the recursive contextual feature with the shape feature extracted for the corresponding part. Our method segments a 3D shape in point cloud into an arbitrary number of parts, depending on the shape complexity, showing strong generality and flexibility. It achieves the state-of-the-art performance, both for fine-grained and semantic segmentation, on the public benchmark and a new benchmark of fine-grained segmentation proposed in this work. We also demonstrate its application for fine-grained part refinements in image-to-shape reconstruction.

****

Learning Multi-Class Segmentations From Single-Class Datasets

Konstantin Dmitriev, Arie E. Kaufman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9501-9511

Multi-class segmentation has recently achieved significant performance in natural images and videos. This achievement is due primarily to the public availability of large multi-class datasets. However, there are certain domains, such as biomedical images, where obtaining sufficient multi-class annotations is a laborious and often impossible task and only single-class datasets are available. While existing segmentation research in such domains use private multi-class datasets or focus on single-class segmentations, we propose a unified highly efficient framework for robust simultaneous learning of multi-class segmentations by combining single-class datasets and utilizing a novel way of conditioning a convolutional network for the purpose of segmentation. We demonstrate various ways of incorporating the conditional information, perform an extensive evaluation, and show compelling multi-class segmentation performance on biomedical images, which outperforms current state-of-the-art solutions (up to 2.7%). Unlike current solutions, which are meticulously tailored for particular single-class datasets, we utilize datasets from a variety of sources. Furthermore, we show the applicability of our method also to natural images and evaluate it on the Cityscapes dataset. We further discuss other possible applications of our proposed framework.

****

Convolutional Recurrent Network for Road Boundary Extraction

Justin Liang, Namdar Homayounfar, Wei-Chiu Ma, Shenlong Wang, Raquel Urtasun ; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit ion (CVPR), 2019, pp. 9512-9521

Creating high definition maps that contain precise information of static element s of the scene is of utmost importance for enabling self driving cars to drive s afely. In this paper, we tackle the problem of drivable road boundary extraction from LiDAR and camera imagery. Towards this goal, we design a structured model where a fully convolutional network obtains deep features encoding the location and direction of road boundaries and then, a convolutional recurrent network out puts a polyline representation for each one of them. Importantly, our method is fully automatic and does not require a user in the loop. We showcase the effect iveness of our method on a large North American city where we obtain perfect to pology of road boundaries 99.3% of the time at a high precision and recall.

****

DFANet: Deep Feature Aggregation for Real-Time Semantic Segmentation

Hanchao Li, Pengfei Xiong, Haoqiang Fan, Jian Sun; Proceedings of the IEEE/CV F Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9522-9 531

This paper introduces an extremely efficient CNN architecture named DFANet for s emantic segmentation under resource constraints. Our proposed network starts fro m a single lightweight backbone and aggregates discriminative features through s ub-network and sub-stage cascade respectively. Based on the multi-scale feature propagation, DFANet substantially reduces the number of parameters, but still ob tains sufficient receptive field and enhances the model learning ability, which strikes a balance between the speed and segmentation performance. Experiments on Cityscapes and CamVid datasets demonstrate the superior performance of DFANet w ith 8xless FLOPs and 2xfaster than the existing state-of-the-art real-time seman tic segmentation methods while providing comparable accuracy. Specifically, it a chieves 70.3% Mean IOU on the Cityscapes test dataset with only 1.7 GFLOPs and a speed of 160 FPS on one NVIDIA Titan X card, and 71.3% Mean IOU with 3.4 GFLOPs while inferring on a higher resolution image.

****

A Cross-Season Correspondence Dataset for Robust Semantic Segmentation

Mans Larsson, Erik Stenborg, Lars Hammarstrand, Marc Pollefeys, Torsten Satt ler, Fredrik Kahl; Proceedings of the IEEE/CVF Conference on Computer Vision an d Pattern Recognition (CVPR), 2019, pp. 9532-9542

In this paper, we present a method to utilize 2D-2D point matches between images taken during different image conditions to train a convolutional neural network for semantic segmentation. Enforcing label consistency across the matches makes the final segmentation algorithm robust to seasonal changes. We describe how th ese 2D-2D matches can be generated with little human interaction by geometricall y matching points from 3D models built from images. Two cross-season corresponde nce datasets are created providing 2D-2D matches across seasonal changes as well as from day to night. The datasets are made publicly available to facilitate fu rther research. We show that adding the correspondences as extra supervision dur ing training improves the segmentation performance of the convolutional neural n etwork, making it more robust to seasonal changes and weather conditions.

****

ManTra-Net: Manipulation Tracing Network for Detection and Localization of Image Forgeries With Anomalous Features

Yue Wu, Wael AbdAlmageed, Premkumar Natarajan; Proceedings of the IEEE/CVF Con ference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9543-9552

To fight against real-life image forgery, which commonly involves different typ es and combined manipulations, we propose a unified deep neural architecture cal led ManTra-Net. Unlike many existing solutions, ManTra-Net is an end-to-end netw ork that performs both detection and localization without extra preprocessing an d postprocessing. \manifold is a fully convolutional network and handles image s of arbitrary sizes and many known forgery types such splicing, copy-move, remo val, enhancement, and even unknown types. This paper has three salient contribut ions. We design a simple yet effective self-supervised learning task to learn ro

bust image manipulation traces from classifying 385 image manipulation types. Fu rther, we formulate the forgery localization problem as a local anomaly detectio n problem, design a Z-score feature to capture local anomaly, and propose a nove l long short-term memory solution to assess local anomalies. Finally, we careful ly conduct ablation experiments to systematically optimize the proposed network design. Our extensive experimental results demonstrate the generalizability, rob ustness and superiority of ManTra-Net, not only in single types of manipulations /forgeries, but also in their complicated combinations.

****

On Zero-Shot Recognition of Generic Objects

Tristan Hascoet, Yasuo Ariki, Tetsuya Takiguchi; Proceedings of the IEEE/CVF C onference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9553-9561

Many recent advances in computer vision are the results of a healthy competition among researchers on high quality, task-specific, benchmarks. After a decade of active research, zero-shot learning (ZSL) models accuracy on the Imagenet bench mark remains far too low to be considered for practical object recognition appli cations. In this paper, we argue that the main reason behind this apparent lack of progress is the poor quality of this benchmark. We highlight major structural flaws of the current benchmark and analyze different factors impacting the accu racy of ZSL models. We show that the actual classification accuracy of existing ZSL models is significantly higher than was previously thought as we account for these flaws. We then introduce the notion of structural bias specific to ZSL da tasets. We discuss how the presence of this new form of bias allows for a trivia l solution to the standard benchmark and conclude on the need for a new benchmar k. We then detail the semi-automated construction of a new benchmark to address these flaws.

****

Explicit Bias Discovery in Visual Question Answering Models

Varun Manjunatha, Nirat Saini, Larry S. Davis; Proceedings of the IEEE/CVF Con ference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9562-9571

Researchers have observed that Visual Question Answering (VQA) models tend to a nswer questions by learning statistical biases in the data. For example, their a nswer to the question "What is the color of the grass?" is usually "Green", wher eas a question like "What is the title of the book?" cannot be answered by infer ring statistical biases. It is of interest to the community to explicitly discov er such biases, both for understanding the behavior of such models, and towards debugging them. Our work address this problem. In a database, we store the words of the question, answer and visual words corresponding to regions of interest i n attention maps. By running simple rule mining algorithms on this database, we discover human-interpretable rules which give us unique insight into the behavio r of such models. Our results also show examples of unusual behaviors learned by models in attempting VQA tasks.

****

REPAIR: Removing Representation Bias by Dataset Resampling

Yi Li, Nuno Vasconcelos; Proceedings of the IEEE/CVF Conference on Computer Vis ion and Pattern Recognition (CVPR), 2019, pp. 9572-9581

Modern machine learning datasets can have biases for certain representations tha t are leveraged by algorithms to achieve high performance without learning to so lve the underlying task. This problem is referred to as "representation bias". T he question of how to reduce the representation biases of a dataset is investiga ted and a new dataset REPresentAtion bIas Removal (REPAIR) procedure is proposed . This formulates bias minimization as an optimization problem, seeking a weight distribution that penalizes examples easy for a classifier built on a given fea ture representation. Bias reduction is then equated to maximizing the ratio betw een the classification loss on the reweighted dataset and the uncertainty of the ground-truth class labels. This is a minimax problem that REPAIR solves by alte rnatingly updating classifier parameters and dataset resampling weights, using s tochastic gradient descent. An experimental set-up is also introduced to measure the bias of any dataset for a given representation, and the impact of this bias on the performance of recognition models. Experiments with synthetic and action

recognition data show that dataset REPAIR can significantly reduce representation bias, and lead to improved generalization of models trained on REPAIRed datasets. The tools used for characterizing representation bias, and the proposed dataset REPAIR algorithm, are available at https://github.com/JerryYLi/Dataset-REPAIR/.

****

Label Efficient Semi-Supervised Learning via Graph Filtering

Qimai Li, Xiao-Ming Wu, Han Liu, Xiaotong Zhang, Zhichao Guan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9582-9591

Graph-based methods have been demonstrated as one of the most effective approaches for semi-supervised learning, as they can exploit the connectivity patterns between labeled and unlabeled data samples to improve learning performance. However, existing graph-based methods either are limited in their ability to jointly model graph structures and data features, such as the classical label propagation methods, or require a considerable amount of labeled data for training and validation due to high model complexity, such as the recent neural-network-based methods. In this paper, we address label efficient semi-supervised learning from a graph filtering perspective. Specifically, we propose a graph filtering framework that injects graph similarity into data features by taking them as signals on the graph and applying a low-pass graph filter to extract useful data representations for classification, where label efficiency can be achieved by conveniently adjusting the strength of the graph filter. Interestingly, this framework unifies two seemingly very different methods -- label propagation and graph convolutional networks. Revisiting them under the graph filtering framework leads to new insights that improve their modeling capabilities and reduce model complexity. Experiments on various semi-supervised classification tasks on four citation networks and one knowledge graph and one semi-supervised regression task for zero-shot image recognition validate our findings and proposals.

****

MVTec AD -- A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection

Paul Bergmann, Michael Fauser, David Sattlegger, Carsten Steger; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9592-9600

The detection of anomalous structures in natural image data is of utmost importance for numerous tasks in the field of computer vision. The development of methods for unsupervised anomaly detection requires data on which to train and evaluate new approaches and ideas. We introduce the MVTec Anomaly Detection (MVTec AD) dataset containing 5354 high-resolution color images of different object and texture categories. It contains normal, i.e., defect-free, images intended for training and images with anomalies intended for testing. The anomalies manifest themselves in the form of over 70 different types of defects such as scratches, dents, contaminations, and various structural changes. In addition, we provide pixel-precise ground truth regions for all anomalies. We also conduct a thorough evaluation of current state-of-the-art unsupervised anomaly detection methods based on deep architectures such as convolutional autoencoders, generative adversarial networks, and feature descriptors using pre-trained convolutional neural networks, as well as classical computer vision methods. This initial benchmark indicates that there is considerable room for improvement. To the best of our knowledge, this is the first comprehensive, multi-object, multi-defect dataset for anomaly detection that provides pixel-accurate ground truth regions and focuses on real-world applications.

****

ABC: A Big CAD Model Dataset for Geometric Deep Learning

Sebastian Koch, Albert Matveev, Zhongshi Jiang, Francis Williams, Alexey Artemov, Evgeny Burnaev, Marc Alexa, Denis Zorin, Daniele Panozzo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9601-9611

We introduce ABC-Dataset, a collection of one million Computer-Aided Design (CAD

) models for research of geometric deep learning methods and applications. Each model is a collection of explicitly parametrized curves and surfaces, providing ground truth for differential quantities, patch segmentation, geometric feature detection, and shape reconstruction. Sampling the parametric descriptions of surfaces and curves allows generating data in different formats and resolutions, enabling fair comparisons for a wide range of geometric learning algorithms. As a use case for our dataset, we perform a large-scale benchmark for estimation of surface normals, comparing existing data driven methods and evaluating their performance against both the ground truth and traditional normal estimation methods.
****

Tightness-Aware Evaluation Protocol for Scene Text Detection
Yuliang Liu, Lianwen Jin, Zecheng Xie, Canjie Luo, Shuaitao Zhang, Lele Xie ; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9612-9620
Evaluation protocols play key role in the developmental progress of text detection methods. There are strict requirements to ensure that the evaluation methods are fair, objective and reasonable. However, existing metrics exhibit some obvious drawbacks: 1) They are not goal-oriented; 2) they cannot recognize the tightness of detection methods; 3) existing one-to-many and many-to-one solutions involve inherent loopholes and deficiencies. Therefore, this paper proposes a novel evaluation protocol called Tightness-aware Intersect-over-Union (TIoU) metric that could quantify completeness of ground truth, compactness of detection, and tightness of matching degree. Specifically, instead of merely using the IoU value, two common detection behaviors are properly considered; meanwhile, directly using the score of TIoU to recognize the tightness. In addition, we further propose a straightforward method to address the annotation granularity issue, which can fairly evaluate word and text-line detections simultaneously. By adopting the detection results from published methods and general object detection frameworks, comprehensive experiments on ICDAR 2013 and ICDAR 2015 datasets are conducted to compare recent metrics and the proposed TIoU metric. The comparison demonstrated some promising new prospects, e.g., determining the methods and frameworks for which the detection is tighter and more beneficial to recognize. Our method is extremely simple; however, the novelty is none other than the proposed metric can utilize simplest but reasonable improvements to lead to many interesting and insightful prospects and solving most the issues of the previous metrics. The code is publicly available at https://github.com/Yuliang-Liu/TIoU-metric.
****

PointConv: Deep Convolutional Networks on 3D Point Clouds
Wenxuan Wu, Zhongang Qi, Li Fuxin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9621-9630
Unlike images which are represented in regular dense grids, 3D point clouds are irregular and unordered, hence applying convolution on them can be difficult. In this paper, we extend the dynamic filter to a new convolution operation, named PointConv. PointConv can be applied on point clouds to build deep convolutional networks. We treat convolution kernels as nonlinear functions of the local coordinates of 3D points comprised of weight and density functions. With respect to a given point, the weight functions are learned with multi-layer perceptron networks and the density functions through kernel density estimation. A novel reformulation is proposed for efficiently computing the weight functions, which allowed us to dramatically scale up the network and significantly improve its performance. The learned convolution kernel can be used to compute translation-invariant and permutation-invariant convolution on any point set in the 3D space. Besides, PointConv can also be used as deconvolution operators to propagate features from a subsampled point cloud back to its original resolution. Experiments on ModelNet40, ShapeNet, and ScanNet show that deep convolutional neural networks built on PointConv are able to achieve state-of-the-art on challenging semantic segmentation benchmarks on 3D point clouds. Besides, our experiments converting CIFAR-10 into a point cloud showed that networks built on PointConv can match the performance of convolutional networks in 2D images of a similar structure.
****

Octree Guided CNN With Spherical Kernels for 3D Point Clouds
Huan Lei, Naveed Akhtar, Ajmal Mian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9631-9640
We propose an octree guided neural network architecture and spherical convolutional kernel for machine learning from arbitrary 3D point clouds. The network architecture capitalizes on the sparse nature of irregular point clouds,and hierarchically coarsens the data representation with space partitioning. At the same time, the proposed spherical kernels systematically quantize point neighborhoods to identify local geometric structures in the data, while maintaining the properties of translation-invariance and asymmetry. We specify spherical kernels with the help of network neurons that in turn are associated with spatial locations.We exploit this association to avert dynamic kernel generation during network training that enables efficient learning with high resolution point clouds. The effectiveness of the proposed technique is established on the benchmark tasks of 3D object classification and segmentation, achieving competitive performance on ShapeNet and RueMonge2014 datasets.
****

VITAMIN-E: VIsual Tracking and MappINg With Extremely Dense Feature Points
Masashi Yokozuka, Shuji Oishi, Simon Thompson, Atsuhiko Banno; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9641-9650
In this paper, we propose a novel indirect monocular simultaneous localization and mapping (SLAM) algorithm called "VITAMIN-E," which is highly accurate and robust as a result of tracking extremely dense feature points. Typical indirect methods have difficulty in reconstructing dense geometry because of their careful feature point selection for accurate matching. Unlike conventional methods, the proposed method processes an enormous number of feature points using the tracking local extrema of curvature based on dominant flow estimation. Because this may lead to high computational cost during bundle adjustment, we propose a novel optimization technique called the "subspace Newton's method" that significantly improves the computational efficiency of bundle adjustment by partially updating the variables. We concurrently generate meshes from the reconstructed points and merge them for an entire three-dimensional(3D) model. Experimental results on the SLAM benchmark EuRoC demonstrated that the proposed method outperformed state-of-the-art SLAM methods such as DSO, ORB-SLAM, and LSD-SLAM, both in terms of accuracy and robustness in trajectory estimation. The proposed method simultaneously generated significantly detailed 3D geometry as a result of the dense feature points in real time using only a CPU.
****

Conditional Single-View Shape Generation for Multi-View Stereo Reconstruction
Yi Wei, Shaohui Liu, Wang Zhao, Jiwen Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9651-9660
In this paper, we present a new perspective towards image-based shape generation. Most existing deep learning based shape reconstruction methods employ a single-view deterministic model which is sometimes insufficient to determine a single groundtruth shape because the back part is occluded. In this work, we first introduce a conditional generative network to model the uncertainty for single-view reconstruction. Then, we formulate the task of multi-view reconstruction as taking the intersection of the predicted shape spaces on each single image. We design new differentiable guidance including the front constraint, the diversity constraint, and the consistency loss to enable effective single-view conditional generation and multi-view synthesis. Experimental results and ablation studies show that our proposed approach outperforms state-of-the-art methods on 3D reconstruction test error and demonstrates its generalization ability on real world data.
****

Learning to Adapt for Stereo
Alessio Tonioni, Oscar Rahnama, Thomas Joy, Luigi Di Stefano, Thalaiyasingam Ajanthan, Philip H.S. Torr; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9661-9670
Real world applications of stereo depth estimation require models that are robus

t to dynamic variations in the environment. Even though deep learning based stereo methods are successful, they often fail to generalize to unseen variations in the environment, making them less suitable for practical applications such as autonomous driving. In this work, we introduce a "learning-to-adapt" framework that enables deep stereo methods to continuously adapt to new target domains in an unsupervised manner. Specifically, our approach incorporates the adaptation procedure into the learning objective to obtain a base set of parameters that are better suited for unsupervised online adaptation. To further improve the quality of the adaptation, we learn a confidence measure that effectively masks the errors introduced during the unsupervised adaptation. We evaluate our method on synthetic and real-world stereo datasets and our experiments evidence that learning-to-adapt is, indeed beneficial for online adaptation on vastly different domains.

****

3D Appearance Super-Resolution With Deep Learning

Yawei Li, Vagia Tsiminaki, Radu Timofte, Marc Pollefeys, Luc Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9671-9680

We tackle the problem of retrieving high-resolution (HR) texture maps of objects that are captured from multiple view points. In the multi-view case, model-based super-resolution (SR) methods have been recently proved to recover high quality texture maps. On the other hand, the advent of deep learning-based methods has already a significant impact on the problem of video and image SR. Yet, a deep learning-based approach to super-resolve the appearance of 3D objects is still missing. The main limitation of exploiting the power of deep learning techniques in the multi-view case is the lack of data. We introduce a 3D appearance SR (3DASR) dataset based on the existing ETH3D [42], SyB3R [31], MiddleBury, and our Collection of 3D scenes from TUM [21], Fountain [51] and Relief [53]. We provide the high- and low-resolution texture maps, the 3D geometric model, images and projection matrices. We exploit the power of 2D learning-based SR methods and design networks suitable for the 3D multi-view case. We incorporate the geometric information by introducing normal maps and further improve the learning process. Experimental results demonstrate that our proposed networks successfully incorporate the 3D geometric information and super-resolve the texture maps.

****

Radial Distortion Triangulation

Zuzana Kukelova, Viktor Larsson; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9681-9689

This paper presents the first optimal, maximal likelihood, solution to the triangulation problem for radially distorted cameras. The proposed solution to the two-view triangulation problem minimizes the L2-norm of the reprojection error in the distorted image space. We cast the problem as the search for corrected distorted image points, and we use a Lagrange multiplier formulation to impose the epipolar constraint for undistorted points. For the one-parameter division model, this formulation leads to a system of five quartic polynomial equations in five unknowns, which can be exactly solved using the Groebner basis method. While the proposed Groebner basis solution is provably optimal; it is too slow for practical applications. Therefore, we developed a fast iterative solver to this problem. Extensive empirical tests show that the iterative algorithm delivers the optimal solution virtually every time, thus making it an L2-optimal algorithm de facto. It is iterative in nature, yet in practice, it converges in no more than five iterations. We thoroughly evaluate the proposed method on both synthetic and real-world data, and we show the benefits of performing the triangulation in the distorted space in the presence of radial distortion.

****

Robust Point Cloud Based Reconstruction of Large-Scale Outdoor Scenes

Ziquan Lan, Zi Jian Yew, Gim Hee Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9690-9698

Outlier feature matches and loop-closures that survived front-end data association can lead to catastrophic failures in the back-end optimization of large-scale

point cloud based 3D reconstruction. To alleviate this problem, we propose a probabilistic approach for robust back-end optimization in the presence of outliers. More specifically, we model the problem as a Bayesian network and solve it using the Expectation-Maximization algorithm. Our approach leverages on a long-tail Cauchy distribution to suppress outlier feature matches in the odometry constraints, and a Cauchy-Uniform mixture model with a set of binary latent variables to simultaneously suppress outlier loop-closure constraints and outlier feature matches in the inlier loop-closure constraints. Furthermore, we show that by using a Gaussian-Uniform mixture model, our approach degenerates to the formulation of a state-of-the-art approach for robust indoor reconstruction. Experimental results demonstrate that our approach has comparable performance with the state-of-the-art on a benchmark indoor dataset, and outperforms it on a large-scale outdoor dataset. Our source code can be found on the project website.

****

Minimal Solvers for Mini-Loop Closures in 3D Multi-Scan Alignment

Pedro Miraldo, Surojit Saha, Srikumar Ramalingam; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9699-9708

3D scan registration is a classical, yet a highly useful problem in the context of 3D sensors such as Kinect and Velodyne. While there are several existing methods, the techniques are usually incremental where adjacent scans are registered first to obtain the initial poses, followed by motion averaging and bundle-adjustment refinement. In this paper, we take a different approach and develop minimal solvers for jointly computing the initial poses of cameras in small loops such as 3-, 4-, and 5-cycles. Note that the classical registration of 2 scans can be done using a minimum of 3 point matches to compute 6 degrees of relative motion. On the other hand, to jointly compute the 3D registrations in n-cycles, we take 2 point matches between the first n-1 consecutive pairs (i.e., Scan 1 & Scan 2, ... , and Scan n-1 & Scan n) and 1 or 2 point matches between Scan 1 and Scan n. Overall, we use 5, 7, and 10 point matches for 3-, 4-, and 5-cycles, and recover 12, 18, and 24 degrees of transformation variables, respectively. Using simulations and real-data we show that the 3D registration using mini n-cycles are computationally efficient, and can provide alternate and better initial poses compared to standard pairwise methods.

****

Volumetric Capture of Humans With a Single RGBD Camera via Semi-Parametric Learning

Rohit Pandey, Anastasia Tkach, Shuoran Yang, Pavel Pidlypenskyi, Jonathan Taylor, Ricardo Martin-Brualla, Andrea Tagliasacchi, George Papandreou, Philip Davidson, Cem Keskin, Shahram Izadi, Sean Fanello; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9709-9718

Volumetric (4D) performance capture is fundamental for AR/VR content generation. Whereas previous work in 4D performance capture has shown impressive results in studio settings, the technology is still far from being accessible to a typical consumer who, at best, might own a single RGBD sensor. Thus, in this work, we propose a method to synthesize free viewpoint renderings using a single RGBD camera. The key insight is to leverage previously seen "calibration" images of a given user to extrapolate what should be rendered in a novel viewpoint from the data available in the sensor. Given these past observations from multiple viewpoints, and the current RGBD image from a fixed view, we propose an end-to-end framework that fuses both these data sources to generate novel renderings of the performer. We demonstrate that the method can produce high fidelity images, and handle extreme changes in subject pose and camera viewpoints. We also show that the system generalizes to performers not seen in the training data. We run exhaustive experiments demonstrating the effectiveness of the proposed semi-parametric model (i.e. calibration images available to the neural network) compared to other state of the art machine learned solutions. Further, we compare the method with more traditional pipelines that employ multi-view capture. We show that our framework is able to achieve compelling results, with substantially less infrastructu

re than previously required.
****
Joint Face Detection and Facial Motion Retargeting for Multiple Faces
Bindita Chaudhuri, Noranart Vesdapunt, Baoyuan Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9719-9728
Facial motion retargeting is an important problem in both computer graphics and vision, which involves capturing the performance of a human face and transferring it to another 3D character. Learning 3D morphable model (3DMM) parameters from 2D face images using convolutional neural networks is common in 2D face alignment, 3D face reconstruction etc. However, existing methods either require an additional face detection step before retargeting or use a cascade of separate networks to perform detection followed by retargeting in a sequence. In this paper, we present a single end-to-end network to jointly predict the bounding box locations and 3DMM parameters for multiple faces. First, we design a novel multitask learning framework that learns a disentangled representation of 3DMM parameters for a single face. Then, we leverage the trained single face model to generate ground truth 3DMM parameters for multiple faces to train another network that performs joint face detection and motion retargeting for images with multiple faces. Experimental results show that our joint detection and retargeting network has high face detection accuracy and is robust to extreme expressions and poses while being faster than state-of-the-art methods.
****
Monocular Depth Estimation Using Relative Depth Maps
Jae-Han Lee, Chang-Su Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9729-9738
We propose a novel algorithm for monocular depth estimation using relative depth maps. First, using a convolutional neural network, we estimate relative depths between pairs of regions, as well as ordinary depths, at various scales. Second, we restore relative depth maps from selectively estimated data based on the rank-1 property of pairwise comparison matrices. Third, we decompose ordinary and relative depth maps into components and recombine them optimally to reconstruct a final depth map. Experimental results show that the proposed algorithm provides the state-of-art depth estimation performance.
****
Unsupervised Primitive Discovery for Improved 3D Generative Modeling
Salman H. Khan, Yulan Guo, Munawar Hayat, Nick Barnes; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9739-9748
3D shape generation is a challenging problem due to the high-dimensional output space and complex part configurations of real-world objects. As a result, existing algorithms experience difficulties in accurate generative modeling of 3D shapes. Here, we propose a novel factorized generative model for 3D shape generation that sequentially transitions from coarse to fine scale shape generation. To this end, we introduce an unsupervised primitive discovery algorithm based on a higher-order conditional random field model. Using the primitive parts for shapes as attributes, a parameterized 3D representation is modeled in the first stage. This representation is further refined in the next stage by adding fine scale details to shape. Our results demonstrate improved representation ability of the generative model and better quality samples of newly generated 3D shapes. Further, our primitive generation approach can accurately parse common objects into a simplified representation.
****
Learning to Explore Intrinsic Saliency for Stereoscopic Video
Qiudan Zhang, Xu Wang, Shiqi Wang, Shikai Li, Sam Kwong, Jianmin Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9749-9758
The human visual system excels at biasing the stereoscopic visual signals by the attention mechanisms. Traditional methods relying on the low-level features and depth relevant information for stereoscopic video saliency prediction have fund

amental limitations. For example, it is cumbersome to model the interactions between multiple visual cues including spatial, temporal, and depth information as a result of the sophistication. In this paper, we argue that the high-level features are crucial and resort to the deep learning framework to learn the saliency map of stereoscopic videos. Driven by spatio-temporal coherence from consecutive frames, the model first imitates the mechanism of saliency by taking advantage of the 3D convolutional neural network. Subsequently, the saliency originated from the intrinsic depth is derived based on the correlations between left and right views in a data-driven manner. Finally, a Convolutional Long Short-Term Memory (Conv-LSTM) based fusion network is developed to model the instantaneous interactions between spatio-temporal and depth attributes, such that the ultimate stereoscopic saliency maps over time are produced. Moreover, we establish a new large-scale stereoscopic video saliency dataset (SVS) including 175 stereoscopic video sequences and their fixation density annotations, aiming to comprehensively study the intrinsic attributes for stereoscopic video saliency detection. Extensive experiments show that our proposed model can achieve superior performance compared to the state-of-the-art methods on the newly built dataset for stereoscopic videos.

****

Spherical Regression: Learning Viewpoints, Surface Normals and 3D Rotations on N-Spheres

Shuai Liao, Efstratios Gavves, Cees G. M. Snoek; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9759-9767

Many computer vision challenges require continuous outputs, but tend to be solved by discrete classification. The reason is classification's natural containment within a probability n-simplex, as defined by the popular softmax activation function. Regular regression lacks such a closed geometry, leading to unstable training and convergence to suboptimal local minima. Starting from this insight we revisit regression in convolutional neural networks. We observe many continuous output problems in computer vision are naturally contained in closed geometrical manifolds, like the Euler angles in viewpoint estimation or the normals in surface normal estimation. A natural framework for posing such continuous output problems are n-spheres, which are naturally closed geometric manifolds defined in the R^(n+1) space. By introducing a spherical exponential mapping on n-spheres at the regression output, we obtain well-behaved gradients, leading to stable training. We show how our spherical regression can be utilized for several computer vision challenges, specifically viewpoint estimation, surface normal estimation and 3D rotation estimation. For all these problems our experiments demonstrate the benefit of spherical regression. All paper resources are available at https://github.com/leoshine/Spherical_Regression.

****

Refine and Distill: Exploiting Cycle-Inconsistency and Knowledge Distillation for Unsupervised Monocular Depth Estimation

Andrea Pilzer, Stephane Lathuiliere, Nicu Sebe, Elisa Ricci; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9768-9777

Nowadays, the majority of state of the art monocular depth estimation techniques are based on supervised deep learning models. However, collecting RGB images with associated depth maps is a very time consuming procedure. Therefore, recent works have proposed deep architectures for addressing the monocular depth prediction task as a reconstruction problem, thus avoiding the need of collecting ground-truth depth. Following these works, we propose a novel self-supervised deep model for estimating depth maps. Our framework exploits two main strategies: refinement via cycle-inconsistency and distillation. Specifically, first a student network is trained to predict a disparity map such as to recover from a frame in a camera view the associated image in the opposite view. Then, a backward cycle network is applied to the generated image to re-synthesize back the input image, estimating the opposite disparity. A third network exploits the inconsistency between the original and the reconstructed input frame in order to output a refined depth map. Finally, knowledge distillation is exploited, such as to transfer i

nformation from the refinement network to the student. Our extensive experimenta l evaluation demonstrate the effectiveness of the proposed framework which outpe rforms state of the art unsupervised methods on the KITTI benchmark.
****

Learning View Priors for Single-View 3D Reconstruction
Hiroharu Kato, Tatsuya Harada; Proceedings of the IEEE/CVF Conference on Comput er Vision and Pattern Recognition (CVPR), 2019, pp. 9778-9787
There is some ambiguity in the 3D shape of an object when the number of observed views is small. Because of this ambiguity, although a 3D object reconstructor c an be trained using a single view or a few views per object, reconstructed shape s only fit the observed views and appear incorrect from the unobserved viewpoint s. To reconstruct shapes that look reasonable from any viewpoint, we propose to train a discriminator that learns prior knowledge regarding possible views. The discriminator is trained to distinguish the reconstructed views of the observed viewpoints from those of the unobserved viewpoints. The reconstructor is trained to correct unobserved views by fooling the discriminator. Our method outperform s current state-of-the-art methods on both synthetic and natural image datasets; this validates the effectiveness of our method.
****

Geometry-Aware Symmetric Domain Adaptation for Monocular Depth Estimation
Shanshan Zhao, Huan Fu, Mingming Gong, Dacheng Tao; Proceedings of the IEEE/C VF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9788-9798
Supervised depth estimation has achieved high accuracy due to the advanced deep network architectures. Since the groundtruth depth labels are hard to obtain, re cent methods try to learn depth estimation networks in an unsupervised way by ex ploring unsupervised cues, which are effective but less reliable than true label s. An emerging way to resolve this dilemma is to transfer knowledge from synthet ic images with ground truth depth via domain adaptation techniques. However, the se approaches overlook specific geometric structure of the natural images in the target domain (i.e., real data), which is important for high-performing depth p rediction. Motivated by the observation, we propose a geometry-aware symmetric d omain adaptation framework (GASDA) to explore the labels in the synthetic data a nd epipolar geometry in the real data jointly. Moreover, by training two image s tyle translators and depth estimators symmetrically in an end-to-end network, ou r model achieves better image style transfer and generates high-quality depth ma ps. The experimental results demonstrate the effectiveness of our proposed metho d and comparable performance against the state-of-the-art.
****

Learning Monocular Depth Estimation Infusing Traditional Stereo Knowledge
Fabio Tosi, Filippo Aleotti, Matteo Poggi, Stefano Mattoccia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9799-9809
Depth estimation from a single image represents a fascinating, yet challenging p roblem with countless applications. Recent works proved that this task could be learned without direct supervision from ground truth labels leveraging image syn thesis on sequences or stereo pairs. Focusing on this second case, in this paper we leverage stereo matching in order to improve monocular depth estimation. To this aim we propose monoResMatch, a novel deep architecture designed to infer de pth from a single input image by synthesizing features from a different point of view, horizontally aligned with the input image, performing stereo matching bet ween the two cues. In contrast to previous works sharing this rationale, our net work is the first trained end-to-end from scratch. Moreover, we show how obtaini ng proxy ground truth annotation through traditional stereo algorithms, such as Semi-Global Matching, enables more accurate monocular depth estimation still cou ntering the need for expensive depth labels by keeping a self-supervised approac h. Exhaustive experimental results prove how the synergy between i) the proposed monoResMatch architecture and ii) proxy-supervision attains state-of-the-art fo r self-supervised monocular depth estimation. The code is publicly available at https://github.com/fabiotosi92/monoResMatch-Tensorflow.

****
SIGNet: Semantic Instance Aided Unsupervised 3D Geometry Perception
Yue Meng, Yongxi Lu, Aman Raj, Samuel Sunarjo, Rui Guo, Tara Javidi, Gaurav Bansal, Dinesh Bharadia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9810-9820
Unsupervised learning for geometric perception (depth, optical flow, etc.) is of great interest to autonomous systems. Recent works on unsupervised learning have made considerable progress on perceiving geometry; however, they usually ignore the coherence of objects and perform poorly under scenarios with dark and noisy environments. In contrast, supervised learning algorithms, which are robust, require large labeled geometric dataset. This paper introduces SIGNet, a novel framework that provides robust geometry perception without requiring geometrically informative labels. Specifically, SIGNet integrates semantic information to make depth and flow predictions consistent with objects and robust to low lighting conditions. SIGNet is shown to improve upon the state-of-the-art unsupervised learning for depth prediction by 30% (in squared relative error). In particular, SIGNet improves the dynamic object class performance by 39% in depth prediction and 29% in flow prediction. Our code will be made available at https://github.com/mengyuest/SIGNet
****
3D Guided Fine-Grained Face Manipulation
Zhenglin Geng, Chen Cao, Sergey Tulyakov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9821-9830
We present a method for fine-grained face manipulation. Given a face image with an arbitrary expression, our method can synthesize another arbitrary expression by the same person. This is achieved by first fitting a 3D face model and then disentangling the face into a texture and a shape. We then learn different networks in these two spaces. In the texture space, we use a conditional generative network to change the appearance, and carefully design input formats and loss functions to achieve the best results. In the shape space, we use a fully connected network to predict the accurate shapes and use the available depth data for supervision. Both networks are conditioned on expression coefficients rather than discrete labels, allowing us to generate an unlimited amount of expressions. We show the superiority of this disentangling approach through both quantitative and qualitative studies. In a user study, our method is preferred in 85% of cases when compared to the most recent work. When compared to the ground truth, annotators cannot reliably distinguish between our synthesized images and real images, preferring our method in 53% of the cases.
****
Neuro-Inspired Eye Tracking With Eye Movement Dynamics
Kang Wang, Hui Su, Qiang Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9831-9840
Generalizing eye tracking to new subjects/environments remains challenging for existing appearance-based methods. To address this issue, we propose to leverage on eye movement dynamics inspired by neurological studies. Studies show that there exist several common eye movement types, independent of viewing contents and subjects, such as fixation, saccade, and smooth pursuits. Incorporating generic eye movement dynamics can therefore improve the generalization capabilities. In particular, we propose a novel Dynamic Gaze Transition Network (DGTN) to capture the underlying eye movement dynamics and serve as the topdown gaze prior. Combined with the bottom-up gaze measurements from the deep convolutional neural network, our method achieves better performance for both within-dataset and cross-dataset evaluations compared to state-of-the-art. In addition, a new DynamicGaze dataset is also constructed to study eye movement dynamics and eye gaze estimation.
****
Facial Emotion Distribution Learning by Exploiting Low-Rank Label Correlations Locally
Xiuyi Jia, Xiang Zheng, Weiwei Li, Changqing Zhang, Zechao Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 201

9, pp. 9841-9850
Emotion recognition from facial expressions is an interesting and challenging pr
oblem and has attracted much attention in recent years. Substantial previous res
earch has only been able to address the ambiguity of "what describes the express
ion", which assumes that each facial expression is associated with one or more p
redefined affective labels while ignoring the fact that multiple emotions always
 have different intensities in a single picture. Therefore, to depict facial exp
ressions more accurately, this paper adopts a label distribution learning approa
ch for emotion recognition that can address the ambiguity of "how to describe th
e expression" and proposes an emotion distribution learning method that exploits
 label correlations locally. Moreover, a local low-rank structure is employed to
 capture the local label correlations implicitly. Experiments on benchmark facia
l expression datasets demonstrate that our method can better address the emotion
 distribution recognition problem than state-of-the-art methods.
****

Unsupervised Face Normalization With Extreme Pose and Expression in the Wild
Yichen Qian, Weihong Deng, Jiani Hu; Proceedings of the IEEE/CVF Conference on
 Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9851-9858
Face recognition achieves great success thanks to the emergence of deep learning
. However, many contemporary face recognition models still have limited invarian
ce to strong intra-personal variations such as large pose changes. Face normaliz
ation provides an effective and cheap way to distil face identity and dispel fac
e variances for recognition. We focus on face generation in the wild with unpair
ed data. To this end, we propose a Face Normalization Model (FNM) to generate a
frontal, neutral expression, photorealistic face image for face recognition. FNM
 is a well-designed Generative Adversarial Network (GAN) with three distinct nov
elties. First, a face expert network is introduced to construct generator and pr
ovide the ability of retaining face identity. Second, with the reconstruction of
 normal face, pixel-wise loss is applied to stabilize optimization process. Thir
d, we present a series of face attention discriminators to refine local textures
. FNM could recover canonical-view, expression-free image and directly improve t
he performance of face recognition model. Extensive qualitative and quantitative
 experiments on both controlled and in-the-wild databases demonstrate the superi
ority of our face normalization method.
****

Semantic Component Decomposition for Face Attribute Manipulation
Ying-Cong Chen, Xiaohui Shen, Zhe Lin, Xin Lu, I-Ming Pao, Jiaya Jia; Proce
edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV
PR), 2019, pp. 9859-9867
Deep neural network-based methods were proposed for face attribute manipulation.
 There still exist, however, two major issues, i.e., insufficient visual quality
 (or resolution) of the results and lack of user control. They limit the applica
bility of existing methods since users may have different editing preference on
facial attributes. In this paper, we address these issues by proposing a semanti
c component model. The model decomposes a facial attribute into multiple semanti
c components, each corresponds to a specific face region. This not only allows f
or user control of edit strength on different parts based on their preference, b
ut also makes it effective to remove unwanted edit effect. Further, each semanti
c component is composed of two fundamental elements, which determine the edit ef
fect and region respectively. This property provides fine interactive control. A
s shown in experiments, our model not only produces high-quality results, but al
so allows effective user interaction.
****

R3 Adversarial Network for Cross Model Face Recognition
Ken Chen, Yichao Wu, Haoyu Qin, Ding Liang, Xuebo Liu, Junjie Yan; Proceedi
ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
, 2019, pp. 9868-9876
In this paper, we raise a new problem, namely cross model face recognition (CMFR
), which has considerable economic and social significance. The core of this pro
blem is to make features extracted from different models comparable. However, th

e diversity, mainly caused by different application scenarios, frequent version updating, and all sorts of service platforms, obstructs interaction among different models and poses a great challenge. To solve this problem, from the perspective of Bayesian modelling, we propose R3 Adversarial Network (R3AN) which consists of three paths: Reconstruction, Representation and Regression. We also introduce adversarial learning into the reconstruction path for better performance. Comprehensive experiments on public datasets demonstrate the feasibility of interaction among different models with the proposed framework. When updating the gallery, R3AN conducts the feature transformation nearly 10 times faster than ResNet-101. Meanwhile, the transformed feature distribution is very close to that of target model, and its error rate is incredibly reduced by approximately 75% compared with a naive transformation model. Furthermore, we show that face feature can be deciphered into original face image roughly by the reconstruction path, which may give valuable hints for improving the original face recognition models.
****

Disentangling Latent Hands for Image Synthesis and Pose Estimation
Linlin Yang, Angela Yao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9877-9886
Hand image synthesis and pose estimation from RGB images are both highly challenging tasks due to the large discrepancy between factors of variation ranging from image background content to camera viewpoint. To better analyze these factors of variation, we propose the use of disentangled representations and a disentangled variational autoencoder (dVAE) that allows for specific sampling and inference of these factors. The derived objective from the variational lower bound as well as the proposed training strategy are highly flexible, allowing us to handle crossmodal encoders and decoders as well as semi-supervised learning scenarios. Experiments show that our dVAE can synthesize highly realistic images of the hand specifiable by both pose and image background content and also estimate 3D hand poses from RGB images with accuracy competitive with state-of-the-art on two public benchmarks.
****

Generating Multiple Hypotheses for 3D Human Pose Estimation With Mixture Density Network
Chen Li, Gim Hee Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9887-9895
3D human pose estimation from a monocular image or 2D joints is an ill-posed problem because of depth ambiguity and occluded joints. We argue that 3D human pose estimation from a monocular input is an inverse problem where multiple feasible solutions can exist. In this paper, we propose a novel approach to generate multiple feasible hypotheses of the 3D pose from 2D joints. In contrast to existing deep learning approaches which minimize a mean square error based on an unimodal Gaussian distribution, our method is able to generate multiple feasible hypotheses of 3D pose based on a multimodal mixture density networks. Our experiments show that the 3D poses estimated by our approach from an input of 2D joints are consistent in 2D reprojections, which supports our argument that multiple solutions exist for the 2D-to-3D inverse problem. Furthermore, we show state-of-the-art performance on the Human3.6M dataset in both best hypothesis and multi-view settings, and we demonstrate the generalization capacity of our model by testing on the MPII and MPI-INF-3DHP datasets. Our code is available at the project website.
****

CrossInfoNet: Multi-Task Information Sharing Based Hand Pose Estimation
Kuo Du, Xiangbo Lin, Yi Sun, Xiaohong Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9896-9905
This paper focuses on the topic of vision based hand pose estimation from single depth map using convolutional neural network (CNN). Our main contributions lie in designing a new pose regression network architecture named CrossInfoNet. The proposed CrossInfoNet decomposes hand pose estimation task into palm pose estimation sub-task and finger pose estimation sub-task, and adopts two-branch crossconnection structure to share the beneficial complementary information between the

sub-tasks. Our work is inspired by multi-task information sharing mechanism, wh
ich has been few discussed in hand pose estimation using depth data in previous
publications. In addition, we propose a heat-map guided feature extraction struc
ture to get better feature maps, and train the complete network end-to-end. The
effectiveness of the proposed CrossInfoNet is evaluated with extensively self-co
mparative experiments and in comparison with state-of-the-art methods on four pu
blic hand pose datasets. The code is available.
****

P2SGrad: Refined Gradients for Optimizing Deep Face Models
Xiao Zhang, Rui Zhao, Junjie Yan, Mengya Gao, Yu Qiao, Xiaogang Wang, Hong
sheng Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
Recognition (CVPR), 2019, pp. 9906-9914
Cosine-based softmax losses significantly improve the performance of deep face r
ecognition networks. However, these losses always include sensitive hyper-parame
ters which can make training process unstable, and it is very tricky to set suit
able hyper parameters for a specific dataset. This paper addresses this challeng
e by directly designing the gradients for training in an adaptive manner. We fir
st investigate and unify previous cosine softmax losses from the perspective of
gradients. This unified view inspires us to propose a novel gradient called P2SG
rad (Probability-to-Similarity Gradient), which leverages a cosine similarity in
stead of classification probability to control the gradients for updating neural
network parameters. P2SGrad is adaptive and hyper-parameter free, which makes t
raining process more efficient and faster. We evaluate our P2SGrad on three face
recognition benchmarks, LFW, MegaFace, and IJB-C. The results show that P2SGrad
is stable in training, robust to noise, and achieves state-of-the-art performan
ce on all the three benchmarks.
****

Action Recognition From Single Timestamp Supervision in Untrimmed Videos
Davide Moltisanti, Sanja Fidler, Dima Damen; Proceedings of the IEEE/CVF Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9915-9924
Recognising actions in videos relies on labelled supervision during training, ty
pically the start and end times of each action instance. This supervision is not
only subjective, but also expensive to acquire. Weak video-level supervision ha
s been successfully exploited for recognition in untrimmed videos, however it is
challenged when the number of different actions in training videos increases. W
e propose a method that is supervised by single timestamps located around each a
ction instance, in untrimmed videos. We replace expensive action bounds with sam
pling distributions initialised from these timestamps. We then use the classifie
r's response to iteratively update the sampling distributions. We demonstrate th
at these distributions converge to the location and extent of discriminative act
ion segments. We evaluate our method on three datasets for fine-grained recognit
ion, with increasing number of different actions per video, and show that single
timestamps offer a reasonable compromise between recognition performance and la
belling effort, performing comparably to full temporal supervision. Our update m
ethod improves top-1 test accuracy by up to 5.4%. across the evaluated datasets.
****

Time-Conditioned Action Anticipation in One Shot
Qiuhong Ke, Mario Fritz, Bernt Schiele; Proceedings of the IEEE/CVF Conference
on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9925-9934
The goal of human action anticipation is to predict future actions. Ideally, in
real-world applications such as video surveillance and self-driving systems, fut
ure actions should not only be predicted with high accuracy but also at arbitrar
y and variable time-horizons ranging from short- to long-term predictions. Curre
nt work mostly focuses on predicting the next action and thus long-term predicti
on is achieved by recursive prediction of each next action, which is both ineffi
cient and accumulates errors. In this paper, we propose a novel time-conditioned
method for efficient and effective long-term action anticipation. There are two
key ingredients to our approach. First, by explicitly conditioning our anticipa
tion network on time allows to efficiently anticipate also long-term actions. An
d second, we propose an attended temporal feature and a time-conditioned skip co

nnection to extract relevant and useful information from observations for effect ive anticipation. We conduct extensive experiments on the large-scale Epic-Kitch en and the 50Salads Datasets. Experimental results show that the proposed method is capable of anticipating future actions at both short-term and long-term, and achieves state-of-the-art performance.
****

Dance With Flow: Two-In-One Stream Action Detection
Jiaojiao Zhao, Cees G. M. Snoek; Proceedings of the IEEE/CVF Conference on Comp uter Vision and Pattern Recognition (CVPR), 2019, pp. 9935-9944
The goal of this paper is to detect the spatio-temporal extent of an action. The two-stream detection network based on RGB and flow provides state-of-the-art ac curacy at the expense of a large model-size and heavy computation. We propose to embed RGB and optical-flow into a single two-in-one stream network with new lay ers. A motion condition layer extracts motion information from flow images, whic h is leveraged by the motion modulation layer to generate transformation paramet ers for modulating the low-level RGB features. The method is easily embedded in existing appearance- or two-stream action detection networks, and trained end-to -end. Experiments demonstrate that leveraging the motion condition to modulate R GB features improves detection accuracy. With only half the computation and para meters of the state-of-the-art two-stream methods, our two-in-one stream still a chieves impressive results on UCF101-24, UCFSports and J-HMDB.
****

Representation Flow for Action Recognition
AJ Piergiovanni, Michael S. Ryoo; Proceedings of the IEEE/CVF Conference on Com puter Vision and Pattern Recognition (CVPR), 2019, pp. 9945-9953
In this paper, we propose a convolutional layer inspired by optical flow algorit hms to learn motion representations. Our representation flow layer is a fully-di fferentiable layer designed to capture the `flow' of any representation channel within a convolutional neural network for action recognition. Its parameters for iterative flow optimization are learned in an end-to-end fashion together with the other CNN model parameters, maximizing the action recognition performance. F urthermore, we newly introduce the concept of learning `flow of flow' representa tions by stacking multiple representation flow layers. We conducted extensive ex perimental evaluations, confirming its advantages over previous recognition mode ls using traditional optical flows in both computational speed and performance. The code is publicly available.
****

LSTA: Long Short-Term Attention for Egocentric Action Recognition
Swathikiran Sudhakaran, Sergio Escalera, Oswald Lanz; Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9954 -9963
Egocentric activity recognition is one of the most challenging tasks in video an alysis. It requires a fine-grained discrimination of small objects and their man ipulation. While some methods base on strong supervision and attention mechanism s, they are either annotation consuming or do not take spatio-temporal patterns into account. In this paper we propose LSTA as a mechanism to focus on features from spatial relevant parts while attention is being tracked smoothly across the video sequence. We demonstrate the effectiveness of LSTA on egocentric activity recognition with an end-to-end trainable two-stream architecture, achieving sta te-of-the-art performance on four standard benchmarks.
****

Learning Actor Relation Graphs for Group Activity Recognition
Jianchao Wu, Limin Wang, Li Wang, Jie Guo, Gangshan Wu; Proceedings of the I EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9964-9974
Modeling relation between actors is important for recognizing group activity in a multi-person scene. This paper aims at learning discriminative relation betwee n actors efficiently using deep models. To this end, we propose to build a flexi ble and efficient \rm Actor Relation Graph (ARG) to simultaneously capture the appearance and position relation between actors. Thanks to the Graph Convolutio

nal Network, the connections in ARG could be automatically learned from group ac
tivity videos in an end-to-end manner, and the inference on ARG could be efficie
ntly performed with standard matrix operations. Furthermore, in practice, we com
e up with two variants to sparsify ARG for more effective modeling in videos: sp
atially localized ARG and temporal randomized ARG. We perform extensive experime
nts on two standard group activity recognition datasets: the Volleyball dataset
and the Collective Activity dataset, where state-of-the-art performance is achie
ved on both datasets. We also visualize the learned actor graphs and relation fe
atures, which demonstrate that the proposed ARG is able to capture the discrimin
ative relation information for group activity recognition.
****

A Structured Model for Action Detection
Yubo Zhang, Pavel Tokmakov, Martial Hebert, Cordelia Schmid; Proceedings of t
he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019,
pp. 9975-9984
A dominant paradigm for learning-based approaches in computer vision is training
generic models, such as ResNet for image recognition, or I3D for video understa
nding, on large datasets and allowing them to discover the optimal representatio
n for the problem at hand. While this is an obviously attractive approach, it is
not applicable in all scenarios. We claim that action detection is one such cha
llenging problem - the models that need to be trained are large, and labeled dat
a is expensive to obtain. To address this limitation, we propose to incorporate
domain knowledge into the structure of the model, simplifying optimization. In p
articular, we augment a standard I3D network with a tracking module to aggregate
long-term motion patterns, and use a graph convolutional network to reason abou
t interactions between actors and objects. Evaluated on the challenging AVA data
set, the proposed approach improves over the I3D baseline by 5.5% mAP and over t
he state-of-the-art by 4.8% mAP.
****

Out-Of-Distribution Detection for Generalized Zero-Shot Action Recognition
Devraj Mandal, Sanath Narayan, Sai Kumar Dwivedi, Vikram Gupta, Shuaib Ahmed
, Fahad Shahbaz Khan, Ling Shao; Proceedings of the IEEE/CVF Conference on Com
puter Vision and Pattern Recognition (CVPR), 2019, pp. 9985-9993
Generalized zero-shot action recognition is a challenging problem, where the tas
k is to recognize new action categories that are unavailable during the training
stage, in addition to the seen action categories. Existing approaches suffer fr
om the inherent bias of the learned classifier towards the seen action categorie
s. As a consequence, unseen category samples are incorrectly classified as belon
ging to one of the seen action categories. In this paper, we set out to tackle t
his issue by arguing for a separate treatment of seen and unseen action categori
es in generalized zero-shot action recognition. We introduce an out-of-distribut
ion detector that determines whether the video features belong to a seen or uns
een action category. To train our out-of-distribution detector, video features f
or unseen action categories are synthesized using generative adversarial network
s trained on seen action category features. To the best of our knowledge, we are
the first to propose an out-of-distribution detector based GZSL framework for a
ction recognition in videos. Experiments are performed on three action recogniti
on datasets: Olympic Sports, HMDB51 and UCF101. For generalized zero-shot action
recognition, our proposed approach outperforms the baseline with absolute gains
(in classification accuracy) of 7.0%, 3.4%, and 4.9%, respectively, on these da
tasets.
****

Object Discovery in Videos as Foreground Motion Clustering
Christopher Xie, Yu Xiang, Zaid Harchaoui, Dieter Fox; Proceedings of the IEE
E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 99
94-10003
We consider the problem of providing dense segmentation masks for object discove
ry in videos. We formulate the object discovery problem as foreground motion clu
stering, where the goal is to cluster foreground pixels in videos into different
objects. We introduce a novel pixel-trajectory recurrent neural network that le

arns feature embeddings of foreground pixel trajectories linked across time. By clustering the pixel trajectories using the learned feature embeddings, our method establishes correspondences between foreground object masks across video frames. To demonstrate the effectiveness of our framework for object discovery, we conduct experiments on commonly used datasets for motion segmentation, where we achieve state-of-the-art performance.

****

Towards Natural and Accurate Future Motion Prediction of Humans and Animals

Zhenguang Liu, Shuang Wu, Shuyuan Jin, Qi Liu, Shijian Lu, Roger Zimmermann, Li Cheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10004-10012

Anticipating the future motions of 3D articulate objects is challenging due to its non-linear and highly stochastic nature. Current approaches typically represent the skeleton of an articulate object as a set of 3D joints, which unfortunately ignores the relationship between joints, and fails to encode fine-grained anatomical constraints. Moreover, conventional recurrent neural networks, such as LSTM and GRU, are employed to model motion contexts, which inherently have difficulties in capturing long-term dependencies. To address these problems, we propose to explicitly encode anatomical constraints by modeling their skeletons with a Lie algebra representation. Importantly, a hierarchical recurrent network structure is developed to simultaneously encodes local contexts of individual frames and global contexts of the sequence. We proceed to explore the applications of our approach to several distinct quantities including human, fish, and mouse. Extensive experiments show that our approach achieves more natural and accurate predictions over state-of-the-art methods.

****

Automatic Face Aging in Videos via Deep Reinforcement Learning

Chi Nhan Duong, Khoa Luu, Kha Gia Quach, Nghia Nguyen, Eric Patterson, Tien D. Bui, Ngan Le; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10013-10022

This paper presents a novel approach for synthesizing automatically age-progressed facial images in video sequences using Deep Reinforcement Learning. The proposed method models facial structures and the longitudinal face-aging process of given subjects coherently across video frames. The approach is optimized using a long-term reward, Reinforcement Learning function with deep feature extraction from Deep Convolutional Neural Network. Unlike previous age-progression methods that are only able to synthesize an aged likeness of a face from a single input image, the proposed approach is capable of age-progressing facial likenesses in videos with consistently synthesized facial features across frames. In addition, the deep reinforcement learning method guarantees preservation of the visual identity of input faces after age-progression. Results on videos of our new collected aging face AGFW-v2 database demonstrate the advantages of the proposed solution in terms of both quality of age-progressed faces, temporal smoothness, and cross-age face verification.

****

Multi-Adversarial Discriminative Deep Domain Generalization for Face Presentation Attack Detection

Rui Shao, Xiangyuan Lan, Jiawei Li, Pong C. Yuen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10023-10031

Face presentation attacks have become an increasingly critical issue in the face recognition community. Many face anti-spoofing methods have been proposed, but they cannot generalize well on "unseen" attacks. This work focuses on improving the generalization ability of face anti-spoofing methods from the perspective of the domain generalization. We propose to learn a generalized feature space via a novel multi-adversarial discriminative deep domain generalization framework. In this framework, a multi-adversarial deep domain generalization is performed under a dual-force triplet-mining constraint. This ensures that the learned feature space is discriminative and shared by multiple source domains, and thus is more generalized to new face presentation attacks. An auxiliary face depth supervis

ion is incorporated to further enhance the generalization ability. Extensive exp eriments on four public datasets validate the effectiveness of the proposed meth od.
****

A Content Transformation Block for Image Style Transfer
Dmytro Kotovenko, Artsiom Sanakoyeu, Pingchuan Ma, Sabine Lang, Bjorn Ommer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti on (CVPR), 2019, pp. 10032-10041
Style transfer has recently received a lot of attention, since it allows to stud y fundamental challenges in image understanding and synthesis. Recent work has s ignificantly improved the representation of color and texture and com- putationa l speed and image resolution. The explicit transformation of image content has, however, been mostly neglected: while artistic style affects formal characterist ics of an image, such as color, shape or texture, it also deforms, adds or remov es content details. This paper explicitly focuses on a content-and style-aware s tylization of a content image. Therefore, we introduce a content transformation module between the encoder and decoder. Moreover, we utilize similar content app earing in photographs and style samples to learn how style alters content detail s and we generalize this to other class details. Additionally, this work present s a novel normalization layer critical for high resolution image synthesis. The robustness and speed of our model enables a video stylization in real-time and h igh definition. We perform extensive qualitative and quantitative evaluations to demonstrate the validity of our approach.
****

BeautyGlow: On-Demand Makeup Transfer Framework With Reversible Generative Netwo rk
Hung-Jen Chen, Ka-Ming Hui, Szu-Yu Wang, Li-Wu Tsao, Hong-Han Shuai, Wen-Hu ang Cheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10042-10050
As makeup has been widely-adopted for beautification, finding suitable makeup by virtual makeup applications becomes popular. Therefore, a recent line of studie s proposes to transfer the makeup from a given reference makeup image to the sou rce non-makeup one. However, it is still challenging due to the massive number o f makeup combinations. To facilitate on-demand makeup transfer, in this work, we propose BeautyGlow that decompose the latent vectors of face images derived fro m the Glow model into makeup and non-makeup latent vectors. Since there is no pa ired dataset, we formulate a new loss function to guide the decomposition. After ward, the non-makeup latent vector of a source image and makeup latent vector of a reference image and are effectively combined and revert back to the image dom ain to derive the results. Experimental results show that the transfer quality o f BeautyGlow is comparable to the state-of-the-art methods, while the unique abi lity to manipulate latent vectors allows BeautyGlow to realize on-demand makeup transfer.
****

Style Transfer by Relaxed Optimal Transport and Self-Similarity
Nicholas Kolkin, Jason Salavon, Gregory Shakhnarovich; Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 100 51-10060
The goal of style transfer algorithms is to render the content of one image usin g the style of another. We propose Style Transfer by Relaxed Optimal Transport a nd Self-Similarity (STROTSS), a new optimization-based style transfer algorithm. We extend our method to allow user specified point-to-point or region-to-region control over visual similarity between the style image and the output. Such gui dance can be used to either achieve a particular visual effect or correct errors made by unconstrained style transfer. In order to quantitatively compare our me thod to prior work, we conduct a large-scale user study designed to assess the s tyle-content tradeoff across settings in style transfer algorithms. Our results indicate that for any desired level of content preservation, our method provides higher quality stylization than prior work.
****

Inserting Videos Into Videos
Donghoon Lee, Tomas Pfister, Ming-Hsuan Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10061-10070
In this paper, we introduce a new problem of manipulating a given video by inserting other videos into it. Our main task is, given an object video and a scene video, to insert the object video at a user-specified location in the scene video so that the resulting video looks realistic. We aim to handle different object motions and complex backgrounds without expensive segmentation annotations. As it is difficult to collect training pairs for this problem, we synthesize fake training pairs that can provide helpful supervisory signals when training a neural network with unpaired real data. The proposed network architecture can take both real and fake pairs as input and perform both supervised and unsupervised training in an adversarial learning scheme. To synthesize a realistic video, the network renders each frame based on the current input and previous frames. Within this framework, we observe that injecting noise into previous frames while generating the current frame stabilizes training. We conduct experiments on real-world videos in object tracking and person re-identification benchmark datasets. Experimental results demonstrate that the proposed algorithm is able to synthesize long sequences of realistic videos with a given object video inserted.
****
Learning Image and Video Compression Through Spatial-Temporal Energy Compaction
Zhengxue Cheng, Heming Sun, Masaru Takeuchi, Jiro Katto; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10071-10080
Compression has been an important research topic for many decades, to produce a significant impact on data transmission and storage. Recent advances have shown a great potential of learning based image and video compression. Inspired from related works, in this paper, we present an image compression architecture using a convolutional autoencoder, and then generalize image compression to video compression, by adding an interpolation loop into both encoder and decoder sides. Our basic idea is to realize spatial-temporal energy compaction in learning image and video compression. Thereby, we propose to add a spatial energy compaction-based penalty into loss function, to achieve higher image compression performance. Furthermore, based on temporal energy distribution, we propose to select the number of frames in one interpolation loop, adapting to the motion characteristics of video contents. Experimental results demonstrate that our proposed image compression outperforms the latest image compression standard with MS-SSIM quality metric, and provides higher performance compared with state-of-the-art learning compression methods at high bit rates, which benefits from our spatial energy compaction approach. Meanwhile, our proposed video compression approach with temporal energy compaction can significantly outperform MPEG-4, and is competitive with commonly used H.264. Both our image and video compression can produce more visually pleasant results than traditional standards.
****
Event-Based High Dynamic Range Image and Very High Frame Rate Video Generation Using Conditional Generative Adversarial Networks
Lin Wang, S. Mohammad Mostafavi I., Yo-Sung Ho, Kuk-Jin Yoon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10081-10090
Event cameras have a lot of advantages over traditional cameras, such as low latency, high temporal resolution, and high dynamic range. However, since the outputs of event cameras are the sequences of asynchronous events over time rather than actual intensity images, existing algorithms could not be directly applied. Therefore, it is demanding to generate intensity images from events for other tasks. In this paper, we unlock the potential of event camera-based conditional generative adversarial networks to create images/videos from an adjustable portion of the event data stream. The stacks of space-time coordinates of events are used as inputs and the network is trained to reproduce images based on the spatio-temporal intensity changes. The usefulness of event cameras to generate high dynamic range (HDR) images even in extreme illumination conditions and also non blur

red images under rapid motion is also shown. In addition, the possibility of gen
erating very high frame rate videos is demonstrated, theoretically up to 1 milli
on frames per second(FPS) since the temporal resolution of event cameras is abou
t 1 microsecond. Proposed methods are evaluated by comparing the results with t
he intensity images captured on the same pixel grid-line of events using online
 available real datasets and synthetic datasets produced by the event camera sim
ulator.
****
Enhancing TripleGAN for Semi-Supervised Conditional Instance Synthesis and Class
ification
Si Wu, Guangchang Deng, Jichang Li, Rui Li, Zhiwen Yu, Hau-San Wong; Procee
dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP
R), 2019, pp. 10091-10100
Learning class-conditional data distributions is crucial for Generative Adversar
ial Networks (GAN) in semi-supervised learning. To improve both instance synthes
is and classification in this setting, we propose an enhanced TripleGAN (Enhance
dTGAN) model in this work. We follow the adversarial training scheme of the orig
inal TripleGAN, but completely re-design the training targets of the generator a
nd classifier. Specifically, we adopt feature-semantics matching to enhance the
generator in learning class-conditional distributions from both the aspects of s
tatistics in the latent space and semantics consistency with respect to the gene
rator and classifier. Since a limited amount of labeled data is not sufficient t
o determine satisfactory decision boundaries, we include two classifiers, and in
corporate collaborative learning into our model to provide better guidance for g
enerator training. The synthesized high-fidelity data can in turn be used for im
proving classifier training. In the experiments, the superior performance of our
 approach on multiple benchmark datasets demonstrates the effectiveness of the m
utual reinforcement between the generator and classifiers in facilitating semi-s
upervised instance synthesis and classification.
****
Capture, Learning, and Synthesis of 3D Speaking Styles
Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, Michael J. Bla
ck; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn
ition (CVPR), 2019, pp. 10101-10111
Audio-driven 3D facial animation has been widely explored, but achieving realist
ic, human-like performance is still unsolved. This is due to the lack of availab
le 3D datasets, models, and standard evaluation metrics. To address this, we int
roduce a unique 4D face dataset with about 29 minutes of 4D scans captured at 60
 fps and synchronized audio from 12 speakers. We then train a neural network on
our dataset that factors identity from facial motion. The learned model, VOCA (V
oice Operated Character Animation) takes any speech signal as input--even speech
 in languages other than English--and realistically animates a wide range of adu
lt faces. Conditioning on subject labels during training allows the model to lea
rn a variety of realistic speaking styles. VOCA also provides animator controls
to alter speaking style, identity-dependent facial shape, and pose (i.e. head, j
aw, and eyeball rotations) during animation. To our knowledge, VOCA is the only
realistic 3D facial animation model that is readily applicable to unseen subject
s without retargeting. This makes VOCA suitable for tasks like in-game video, vi
rtual reality avatars, or any scenario in which the speaker, speech, or language
 is not known in advance. We make the dataset and model available for research p
urposes at http://voca.is.tue.mpg.de.
****
Nesti-Net: Normal Estimation for Unstructured 3D Point Clouds Using Convolutiona
l Neural Networks
Yizhak Ben-Shabat, Michael Lindenbaum, Anath Fischer; Proceedings of the IEEE/
CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1011
2-10120
In this paper, we propose a normal estimation method for unstructured 3D point c
louds. This method, called Nesti-Net, builds on a new local point cloud represen
tation which consists of multi-scale point statistics (MuPS), estimated on a loc

al coarse Gaussian grid. This representation is a suitable input to a CNN architecture. The normals are estimated using a mixture-of-experts (MoE) architecture, which relies on a data-driven approach for selecting the optimal scale around each point and encourages sub-network specialization. Interesting insights into the network's resource distribution are provided. The scale prediction significantly improves robustness to different noise levels, point density variations and different levels of detail. We achieve state-of-the-art results on a benchmark synthetic dataset and present qualitative results on real scanned scenes.
****

Ray-Space Projection Model for Light Field Camera
Qi Zhang, Jinbo Ling, Qing Wang, Jingyi Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10121-10129
Light field essentially represents the collection of rays in space. The rays captured by multiple light field cameras form subsets of full rays in 3D space and can be transformed to each other. However, most previous approaches model the projection from an arbitrary point in 3D space to corresponding pixel on the sensor. There are few models on describing the ray sampling and transformation among multiple light field cameras. In the paper, we propose a novel ray-space projection model to transform sets of rays captured by multiple light field cameras in term of the Plucker coordinates. We first derive a 6x6 ray-space intrinsic matrix based on multi-projection-center (MPC) model. A homogeneous ray-space projection matrix and a fundamental matrix are then proposed to establish ray-ray correspondences among multiple light fields. Finally, based on the ray-space projection matrix, a novel camera calibration method is proposed to verify the proposed model. A linear constraint and a ray-ray cost function are established for linear initial solution and non-linear optimization respectively. Experimental results on both synthetic and real light field data have verified the effectiveness and robustness of the proposed model.
****

Deep Geometric Prior for Surface Reconstruction
Francis Williams, Teseo Schneider, Claudio Silva, Denis Zorin, Joan Bruna, Daniele Panozzo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10130-10139
The reconstruction of a discrete surface from a point cloud is a fundamental geometry processing problem that has been studied for decades, with many methods developed. We propose the use of a deep neural network as a geometric prior for surface reconstruction. Specifically, we overfit a neural network representing a local chart parameterization to part of an input point cloud using the Wasserstein distance as a measure of approximation. By jointly fitting many such networks to overlapping parts of the point cloud, while enforcing a consistency condition, we compute a manifold atlas. By sampling this atlas, we can produce a dense reconstruction of the surface approximating the input cloud. The entire procedure does not require any training data or explicit regularization, yet, we show that it is able to perform remarkably well: not introducing typical overfitting artifacts, and approximating sharp features closely at the same time. We experimentally show that this geometric prior produces good results for both man-made objects containing sharp features and smoother organic objects, as well as noisy inputs. We compare our method with a number of well-known reconstruction methods on a standard surface reconstruction benchmark.
****

Analysis of Feature Visibility in Non-Line-Of-Sight Measurements
Xiaochun Liu, Sebastian Bauer, Andreas Velten; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10140-10148
We formulate an equation describing a general Non-line-of-sight (NLOS) imaging measurement and analyze the properties of the measurement in the Fourier domain regarding the spatial frequencies of the scene it encodes. We conclude that for a relay wall with finite size, certain scene configurations and features are not detectable in an NLOS measurement. We then provide experimental examples of invisible scene features and their reconstructions, as well as a set of example scenes that lead to an ill-posed NLOS imaging problem.

****
Hyperspectral Imaging With Random Printed Mask
Yuanyuan Zhao, Hui Guo, Zhan Ma, Xun Cao, Tao Yue, Xuemei Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10149-10157
Hyperspectral images can provide rich clues for various computer vision tasks. However, the requirements of professional and expensive hardware for capturing hyperspectral images impede its wide applications. In this paper, based on a simple but not widely noticed phenomenon that the color printer can print color masks with a large number of independent spectral transmission responses, we propose a simple and low-budget scheme to capture the hyperspectral images with a random mask printed by the consumer-level color printer. Specifically, we notice that the printed dots with different colors are stacked together, forming multiplicative, instead of additive, spectral transmission responses. Therefore, new spectral transmission response uncorrelated with that of the original printer dyes are generated. With the random printed color mask, hyperspectral images could be captured in a snapshot way. A convolutional neural network (CNN) based method is developed to reconstruct the hyperspectral images from the captured image. The effectiveness and accuracy of the proposed system are verified on both synthetic and real captured images.
****
All-Weather Deep Outdoor Lighting Estimation
Jinsong Zhang, Kalyan Sunkavalli, Yannick Hold-Geoffroy, Sunil Hadap, Jonathan Eisenman, Jean-Francois Lalonde; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10158-10166
We present a neural network that predicts HDR outdoor illumination from a single LDR image. At the heart of our work is a method to accurately learn HDR lighting from LDR panoramas under any weather condition. We achieve this by training another CNN (on a combination of synthetic and real images) to take as input an LDR panorama, and regress the parameters of the Lalonde-Mathews outdoor illumination model. This model is trained such that it a) reconstructs the appearance of the sky, and b) renders the appearance of objects lit by this illumination. We use this network to label a large-scale dataset of LDR panoramas with lighting parameters and use them to train our single image outdoor lighting estimation network. We demonstrate, via extensive experiments, that both our panorama and single image networks outperform the state of the art, and unlike prior work, are able to handle weather conditions ranging from fully sunny to overcast skies.
****
A Variational EM Framework With Adaptive Edge Selection for Blind Motion Deblurring
Liuge Yang, Hui Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10167-10176
Blind motion deblurring is an important problem that receives enduring attention in last decade. Based on the observation that a good intermediate estimate of latent image for estimating motion-blur kernel is not necessarily the one closest to latent image, edge selection has proven itself a very powerful technique for achieving state-of-the-art performance in blind deblurring. This paper presented an interpretation of edge selection/reweighting in terms of variational Bayes inference, and therefore developed a novel variational expectation maximization (VEM) algorithm with built-in adaptive edge selection for blind deblurring. Together with a restart strategy for avoiding undesired local convergence, the proposed VEM method not only has a solid mathematical foundation but also noticeably outperformed the state-of-the-art methods on benchmark datasets.
****
Viewport Proposal CNN for 360deg Video Quality Assessment
Chen Li, Mai Xu, Lai Jiang, Shanyi Zhang, Xiaoming Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10177-10186
Recent years have witnessed the growing interest in visual quality assessment (VQA) for 360deg video. Unfortunately, the existing VQA approaches do not consider

the facts that: 1) Observers only see viewports of 360deg video, rather than patches or whole 360deg frames. 2) Within the viewport, only salient regions can be perceived by observers with high resolution. Thus, this paper proposes a viewport-based convolutional neural network (V-CNN) approach for VQA on 360deg video, considering both auxiliary tasks of viewport proposal and viewport saliency prediction. Our V-CNN approach is composed of two stages, i.e., viewport proposal and VQA. In the first stage, the viewport proposal network (VP-net) is developed to yield several potential viewports, seen as the first auxiliary task. In the second stage, a viewport quality network (VQ-net) is designed to rate the VQA score for each proposed viewport, in which the saliency map of the viewport is predicted and then utilized in VQA score rating. Consequently, another auxiliary task of viewport saliency prediction can be achieved. More importantly, the main task of VQA on 360deg video can be accomplished via integrating the VQA scores of all viewports. The experiments validate the effectiveness of our V-CNN approach in significantly advancing the state-of-the-art performance of VQA on 360deg video. In addition, our approach achieves comparable performance in two auxiliary tasks. The code of our V-CNN approach is available at https://github.com/Archer-Tatsu/V-CNN.

****

Beyond Gradient Descent for Regularized Segmentation Losses

Dmitrii Marin, Meng Tang, Ismail Ben Ayed, Yuri Boykov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10187-10196

The simplicity of gradient descent (GD) made it the default method for training ever-deeper and complex neural networks. Both loss functions and architectures are often explicitly tuned to be amenable to this basic local optimization. In the context of weakly-supervised CNN segmentation, we demonstrate a well-motivated loss function where an alternative optimizer (ADM) achieves the state-of-the-art while GD performs poorly. Interestingly, GD obtains its best result for a "smoother" tuning of the loss function. The results are consistent across different network architectures. Our loss is motivated by well-understood MRF/CRF regularization models in "shallow" segmentation and their known global solvers. Our work suggests that network design/training should pay more attention to optimization methods.

****

MAGSAC: Marginalizing Sample Consensus

Daniel Barath, Jiri Matas, Jana Noskova; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10197-10205

A method called, sigma-consensus, is proposed to eliminate the need for a user-defined inlier-outlier threshold in RANSAC. Instead of estimating the noise sigma, it is marginalized over a range of noise scales. The optimized model is obtained by weighted least-squares fitting where the weights come from the marginalization over sigma of the point likelihoods of being inliers. A new quality function is proposed not requiring sigma and, thus, a set of inliers to determine the model quality. Also, a new termination criterion for RANSAC is built on the proposed marginalization approach. Applying sigma-consensus, MAGSAC is proposed with no need for a user-defined sigma and improving the accuracy of robust estimation significantly. It is superior to the state-of-the-art in terms of geometric accuracy on publicly available real-world datasets for epipolar geometry (F and E) and homography estimation. In addition, applying sigma-consensus only once as a post-processing step to the RANSAC output always improved the model quality on a wide range of vision problems without noticeable deterioration in processing time, adding a few milliseconds.

****

Understanding and Visualizing Deep Visual Saliency Models

Sen He, Hamed R. Tavakoli, Ali Borji, Yang Mi, Nicolas Pugeault; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10206-10215

Recently, data-driven deep saliency models have achieved high performance and have outperformed classical saliency models, as demonstrated by results on dataset

s such as the MIT300 and SALICON. Yet, there remains a large gap between the performance of these models and the inter-human baseline. Some outstanding questions include what have these models learned, how and where they fail, and how they can be improved. This article attempts to answer these questions by analyzing the representations learned by individual neurons located at the intermediate layers of deep saliency models. To this end, we follow the steps of existing deep saliency models, that is borrowing a pre-trained model of object recognition to encode the visual features and learning a decoder to infer the saliency. We consider two cases when the encoder is used as a fixed feature extractor and when it is fine-tuned, and compare the inner representations of the network. To study how the learned representations depend on the task, we fine-tune the same network using the same image set but for two different tasks: saliency prediction versus scene classification. Our analyses reveal that: 1) some visual regions (e.g. head, text, symbol, vehicle) are already encoded within various layers of the network pre-trained for object recognition, 2) using modern datasets, we find that fine-tuning pre-trained models for saliency prediction makes them favor some categories (e.g. head) over some others (e.g. text), 3) although deep models of saliency outperform classical models on natural images, the converse is true for synthetic stimuli (e.g. pop-out search arrays), an evidence of significant difference between human and data-driven saliency models, and 4) we confirm that, after-fine tuning, the change in inner-representations is mostly due to the task and not the domain shift in the data
****

Divergence Prior and Vessel-Tree Reconstruction
Zhongwen Zhang, Dmitrii Marin, Egor Chesakov, Marc Moreno Maza, Maria Drangova, Yuri Boykov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10216-10224
We propose a new geometric regularization principle for reconstructing vector fields based on prior knowledge about their divergence. As one important example of this general idea, we focus on vector fields modelling blood flow pattern that should be divergent in arteries and convergent in veins. We show that this previously ignored regularization constraint can significantly improve the quality of vessel tree reconstruction particularly around bifurcations where non-zero divergence is concentrated. Our divergence prior is critical for resolving (binary) sign ambiguity in flow orientations produced by standard vessel filters, e.g. Frangi. Our vessel tree centerline reconstruction combines divergence constraints with robust curvature regularization. Our unsupervised method can reconstruct complete vessel trees with near-capillary details on synthetic and real 3D volumes.
****

Unsupervised Domain-Specific Deblurring via Disentangled Representations
Boyu Lu, Jun-Cheng Chen, Rama Chellappa; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10225-10234
Image deblurring aims to restore the latent sharp images from the corresponding blurred ones. In this paper, we present an unsupervised method for domain-specific, single-image deblurring based on disentangled representations. The disentanglement is achieved by splitting the content and blur features in a blurred image using content encoders and blur encoders. We enforce a KL divergence loss to regularize the distribution range of extracted blur attributes such that little content information is contained. Meanwhile, to handle the unpaired training data, a blurring branch and the cycle-consistency loss are added to guarantee that the content structures of the deblurred results match the original images. We also add an adversarial loss on deblurred results to generate visually realistic images and a perceptual loss to further mitigate the artifacts. We perform extensive experiments on the tasks of face and text deblurring using both synthetic datasets and real images, and achieve improved results compared to recent state-of-the-art deblurring methods.
****

Douglas-Rachford Networks: Learning Both the Image Prior and Data Fidelity Terms for Blind Image Deconvolution

Raied Aljadaany, Dipan K. Pal, Marios Savvides; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10235-10244

Blind deconvolution problems are heavily ill-posed where the specific blurring kernel is not known. Recovering these images typically requires estimates of the kernel. In this paper, we present a method called Dr-Net, which does not require any such estimate and is further able to invert the effects of the blurring in blind image recovery tasks. These image recovery problems typically have two terms, the data fidelity term (for faithful reconstruction) and the image prior (for realistic looking reconstructions). We use the Douglas-Rachford iterations to solve this problem since it is a more generally applicable optimization procedure than methods such as the proximal gradient descent algorithm. Two proximal operators originate from these iterations, one from the data fidelity term and the second from the image prior. It is non-trivial to design a hand-crafted function to represent these proximal operators for the data fidelity and the image prior terms which would work with real-world image distributions. We therefore approximate both these proximal operators using deep networks. This provides a sound motivation for the final architecture for Dr-Net which we find outperforms the state-of-the-art on two mainstream blind deconvolution benchmarks. We also find that Dr-Net is one of the fastest algorithms according to wall-clock times while doing so.

****

Speed Invariant Time Surface for Learning to Detect Corner Points With Event-Based Cameras

Jacques Manderscheid, Amos Sironi, Nicolas Bourdis, Davide Migliore, Vincent Lepetit; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10245-10254

We propose a learning approach to corner detection for event-based cameras that is stable even under fast and abrupt motions. Event-based cameras offer high temporal resolution, power efficiency, and high dynamic range. However, the properties of event-based data are very different compared to standard intensity images, and simple extensions of corner detection methods designed for these images do not perform well on event-based data. We first introduce an efficient way to compute a time surface that is invariant to the speed of the objects. We then show that we can train a Random Forest to recognize events generated by a moving corner from our time surface. Random Forests are also extremely efficient, and therefore a good choice to deal with the high capture frequency of event-based cameras ---our implementation processes up to 1.6 Mev/s on a single CPU. Thanks to our time surface formulation and this learning approach, our method is significantly more robust to abrupt changes of direction of the corners compared to previous ones. Our method also naturally assigns a confidence score for the corners, which can be useful for posttprocessing. Moreover, we introduce a high-resolution dataset suitable for quantitative evaluation and comparison of corner detection methods for event-based cameras. We call our approach SILC, for Speed Invariant Learned Corners, and compare it to the state-of-the-art with extensive experiments, showing better performance.

****

Training Deep Learning Based Image Denoisers From Undersampled Measurements Without Ground Truth and Without Image Prior

Magauiya Zhussip, Shakarim Soltanayev, Se Young Chun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10255-10264

Compressive sensing is a method to recover the original image from undersampled measurements. In order to overcome the ill-posedness of this inverse problem, image priors are used such as sparsity, minimal total-variation, or self-similarity of images. Recently, deep learning based compressive image recovery methods have been proposed and have yielded state-of-the-art performances. They used data-driven approaches instead of hand-crafted image priors to regularize ill-posed inverse problems with undersampled data. Ironically, training deep neural network

s (DNNs) for them requires "clean" ground truth images, but obtaining the best q
uality images from undersampled data requires well-trained DNNs. To resolve this
 dilemma, we propose novel methods based on two well-grounded theories: denoiser
-approximate message passing (D-AMP) and Stein's unbiased risk estimator (SURE).
 Our proposed methods were able to train deep learning based image denoisers fro
m undersampled measurements without ground truth images and without additional i
mage priors, and to recover images with state-of-the-art qualities from undersam
pled data. We evaluated our methods for various compressive sensing recovery pro
blems with Gaussian random, coded diffraction pattern, and compressive sensing M
RI measurement matrices. Our proposed methods yielded state-of-the-art performan
ces for all cases without ground truth images. Our methods also yielded comparab
le performances to the methods with ground truth data.
****

A Variational Pan-Sharpening With Local Gradient Constraints
Xueyang Fu, Zihuang Lin, Yue Huang, Xinghao Ding; Proceedings of the IEEE/CVF
 Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10265-1
0274
Pan-sharpening aims at fusing spectral and spatial information, which are respec
tively contained in the multispectral (MS) image and panchromatic (PAN) image, t
o produce a high resolution multi-spectral (HRMS) image. In this paper, a new va
riational model based on a local gradient constraint for pan-sharpening is propo
sed. Different with previous methods that only use global constraints to preserv
e spatial information, we first consider gradient difference of PAN and HRMS ima
ges in different local patches and bands. Then a more accurate spatial preservat
ion based on local gradient constraints is incorporated into the objective to fu
lly utilize spatial information contained in the PAN image. The objective is for
mulated as a convex optimization problem which minimizes two leastsquares terms
and thus very simple and easy to implement. A fast algorithm is also designed to
 improve efficiency. Experiments show that our method outperforms previous varia
tional algorithms and achieves better generalization than recent deep learning m
ethods.
****

F-VAEGAN-D2: A Feature Generating Framework for Any-Shot Learning
Yongqin Xian, Saurabh Sharma, Bernt Schiele, Zeynep Akata; Proceedings of the
 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp
. 10275-10284
When labeled training data is scarce, a promising data augmentation approach is
to generate visual features of unknown classes using their attributes. To learn
the class conditional distribution of CNN features, these models rely on pairs o
f image features and class attributes. Hence, they can not make use of the abund
ance of unlabeled data samples. In this paper, we tackle any-shot learning probl
ems i.e. zero-shot and few-shot, in a unified feature generating framework that
operates in both inductive and transductive learning settings. We develop a cond
itional generative model that combines the strength of VAE and GANs and in addit
ion, via an unconditional discriminator, learns the marginal feature distributio
n of unlabeled images. We empirically show that our model learns highly discrimi
native CNN features for five datasets, i.e. CUB, SUN, AWA and ImageNet, and esta
blish a new state-of-the-art in any-shot learning, i.e. inductive and transducti
ve (generalized) zero- and few-shot learning settings. We also demonstrate that
our learned features are interpretable: we visualize them by inverting them back
 to the pixel space and we explain them by generating textual arguments of why t
hey are associated with a certain label.
****

Sliced Wasserstein Discrepancy for Unsupervised Domain Adaptation
Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, Daniel Ulbricht; Proceedings
of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20
19, pp. 10285-10295
In this work, we connect two distinct concepts for unsupervised domain adaptatio
n: feature distribution alignment between domains by utilizing the task-specific
 decision boundary and the Wasserstein metric. Our proposed sliced Wasserstein d

iscrepancy (SWD) is designed to capture the natural notion of dissimilarity betw
een the outputs of task-specific classifiers. It provides a geometrically meanin
gful guidance to detect target samples that are far from the support of the sour
ce and enables efficient distribution alignment in an end-to-end trainable fashi
on. In the experiments, we validate the effectiveness and genericness of our met
hod on digit and sign recognition, image classification, semantic segmentation,
and object detection.
****

Graph Attention Convolution for Point Cloud Semantic Segmentation
Lei Wang, Yuchun Huang, Yaolin Hou, Shenman Zhang, Jie Shan; Proceedings of
the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019,
pp. 10296-10305
Standard convolution is inherently limited for semantic segmentation of point cl
oud due to its isotropy about features. It neglects the structure of an object,
results in poor object delineation and small spurious regions in the segmentatio
n result. This paper proposes a novel graph attention convolution (GAC), whose k
ernels can be dynamically carved into specific shapes to adapt to the structure
of an object. Specifically, by assigning proper attentional weights to different
neighboring points, GAC is designed to selectively focus on the most relevant p
art of them according to their dynamically learned features. The shape of the co
nvolution kernel is then determined by the learned distribution of the attention
al weights. Though simple, GAC can capture the structured features of point clou
ds for fine-grained segmentation and avoid feature contamination between objects
. Theoretically, we provided a thorough analysis on the expressive capabilities
of GAC to show how it can learn about the features of point clouds. Empirically,
we evaluated the proposed GAC on challenging indoor and outdoor datasets and ac
hieved the state-of-the-art results in both scenarios.
****

Normalized Diversification
Shaohui Liu, Xiao Zhang, Jianqiao Wangni, Jianbo Shi; Proceedings of the IEEE
/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 103
06-10315
Generating diverse yet specific data is the goal of the generative adversarial n
etwork (GAN), but it suffers from the problem of mode collapse. We introduce the
concept of normalized diversity which force the model to preserve the normalize
d pairwise distance between the sparse samples from a latent parametric distribu
tion and their corresponding high-dimensional outputs. The normalized diversific
ation aims to unfold the manifold of unknown topology and non-uniform distributi
on, which leads to safe interpolation between valid latent variables. By alterna
ting the maximization over the pairwise distance and updating the total distance
(normalizer), we encourage the model to actively explore in the high-dimensiona
l output space. We demonstrate that by combining the normalized diversity loss a
nd the adversarial loss, we generate diverse data without suffering from mode co
llapsing. Experimental results show that our method achieves consistent improvem
ent on unsupervised image generation, conditional image generation and hand pose
estimation over strong baselines.
****

Learning to Localize Through Compressed Binary Maps
Xinkai Wei, Ioan Andrei Barsan, Shenlong Wang, Julieta Martinez, Raquel Urta
sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog
nition (CVPR), 2019, pp. 10316-10324
One of the main difficulties of scaling current localization systems to large en
vironments is the on-board storage required for the maps. In this paper we propo
se to learn to compress the map representation such that it is optimal for the l
ocalization task. As a consequence, higher compression rates can be achieved wit
hout loss of localization accuracy when compared to standard coding schemes that
optimize for reconstruction, thus ignoring the end task. Our experiments show t
hat it is possible to learn a task-specific compression which reduces storage re
quirements by two orders of magnitude over general-purpose codecs such as WebP w
ithout sacrificing performance.

****
A Parametric Top-View Representation of Complex Road Scenes
Ziyan Wang, Buyu Liu, Samuel Schulter, Manmohan Chandraker; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10325-10333
In this paper, we address the problem of inferring the layout of complex road scenes given a single camera as input. To achieve that, we first propose a novel parameterized model of road layouts in a top-view representation, which is not only intuitive for human visualization but also provides an interpretable interface for higher-level decision making. Moreover, the design of our top-view scene model allows for efficient sampling and thus generation of large-scale simulated data, which we leverage to train a deep neural network to infer our scene model's parameters. Specifically, our proposed training procedure uses supervised domain-adaptation techniques to incorporate both simulated as well as manually annotated data. Finally, we design a Conditional Random Field (CRF) that enforces coherent predictions for a single frame and encourages temporal smoothness among video frames. Experiments on two public data sets show that: (1) Our parametric top-view model is representative enough to describe complex road scenes; (2) The proposed method outperforms baselines trained on manually-annotated or simulated data only, thus getting the best of both; (3) Our CRF is able to generate temporally smoothed while semantically meaningful results.
****
Self-Supervised Spatiotemporal Learning via Video Clip Order Prediction
Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, Yueting Zhuang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10334-10343
We propose a self-supervised spatiotemporal learning technique which leverages the chronological order of videos. Our method can learn the spatiotemporal representation of the video by predicting the order of shuffled clips from the video. The category of the video is not required, which gives our technique the potential to take advantage of infinite unannotated videos. There exist related works which use frames, while compared to frames, clips are more consistent with the video dynamics. Clips can help to reduce the uncertainty of orders and are more appropriate to learn a video representation. The 3D convolutional neural networks are utilized to extract features for clips, and these features are processed to predict the actual order. The learned representations are evaluated via nearest neighbor retrieval experiments. We also use the learned networks as the pre-trained models and finetune them on the action recognition task. Three types of 3D convolutional neural networks are tested in experiments, and we gain large improvements compared to existing self-supervised methods.
****
Superquadrics Revisited: Learning 3D Shape Parsing Beyond Cuboids
Despoina Paschalidou, Ali Osman Ulusoy, Andreas Geiger; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10344-10353
Abstracting complex 3D shapes with parsimonious part-based representations has been a long standing goal in computer vision. This paper presents a learning-based solution to this problem which goes beyond the traditional 3D cuboid representation by exploiting superquadrics as atomic elements. We demonstrate that superquadrics lead to more expressive 3D scene parses while being easier to learn than 3D cuboid representations. Moreover, we provide an analytical solution to the Chamfer loss which avoids the need for computational expensive reinforcement learning or iterative prediction. Our model learns to parse 3D objects into consistent superquadric representations without supervision. Results on various ShapeNet categories as well as the SURREAL human body dataset demonstrate the flexibility of our model in capturing fine details and complex poses that could not have been modelled using cuboids.
****
Unsupervised Disentangling of Appearance and Geometry by Deformable Generator Network

Xianglei Xing, Tian Han, Ruiqi Gao, Song-Chun Zhu, Ying Nian Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10354-10363

We present a deformable generator model to disentangle the appearance and geometric information in purely unsupervised manner. The appearance generator models the appearance related information, including color, illumination, identity or category, of an image, while the geometric generator performs geometric related warping, such as rotation and stretching, through generating displacement of the coordinates of each pixel to obtain the final image. Two generators act upon independent latent factors to extract disentangled appearance and geometric information from image. The proposed scheme is general and can be easily integrated into different generative models. An extensive set of qualitative and quantitative experiments show that the appearance and geometric information can be well disentangled, and the learned geometric generator can be conveniently transferred to the other image datasets to facilitate knowledge transfer tasks.

****

Self-Supervised Representation Learning by Rotation Feature Decoupling

Zeyu Feng, Chang Xu, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10364-10374

We introduce a self-supervised learning method that focuses on beneficial properties of representation and their abilities in generalizing to real-world tasks. The method incorporates rotation invariance into the feature learning framework, one of many good and well-studied properties of visual representation, which is rarely appreciated or exploited by previous deep convolutional neural network based self-supervised representation learning methods. Specifically, our model learns a split representation that contains both rotation related and unrelated parts. We train neural networks by jointly predicting image rotations and discriminating individual instances. In particular, our model decouples the rotation discrimination from instance discrimination, which allows us to improve the rotation prediction by mitigating the influence of rotation label noise, as well as discriminate instances without regard to image rotations. The resulting feature has a better generalization ability for more various tasks. Experimental results show that our model outperforms current state-of-the-art methods on standard self-supervised feature learning benchmarks.

****

Weakly Supervised Deep Image Hashing Through Tag Embeddings

Vijetha Gattupalli, Yaoxin Zhuo, Baoxin Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10375-10384

Many approaches to semantic image hashing have been formulated as supervised learning problems that utilize images and label information to learn the binary hash codes. However, large-scale labelled image data is expensive to obtain, thus imposing a restriction on the usage of such algorithms. On the other hand, unlabelled image data is abundant due to the existence of many Web image repositories. Such Web images may often come with images tags that contains useful information, although raw tags in general do not readily lead to semantic labels. Motivated by this scenario, we formulate the problem of semantic image hashing as a weakly-supervised learning problem. We utilize the information contained in the user-generated tags associated with the images to learn the hash codes. More specifically, we extract the word2vec semantic embeddings of the tags and use the information contained in them for constraining the learning. Accordingly, we name our model Weakly Supervised Deep Hashing using Tag Embeddings (WDHT). WDHT is tested for the task of semantic image retrieval and is compared against several state-of-art models. Results show that our approach sets a new state-of-art in the area of weekly supervised image hashing.

****

Improved Road Connectivity by Joint Learning of Orientation and Segmentation

Anil Batra, Suriya Singh, Guan Pang, Saikat Basu, C.V. Jawahar, Manohar Paluri; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10385-10393

Road network extraction from satellite images often produce fragmented road segm

ents leading to road maps unfit for real applications. Pixel-wise classification fails to predict topologically correct and connected road masks due to the absence of connectivity supervision and difficulty in enforcing topological constraints. In this paper, we propose a connectivity task called Orientation Learning, motivated by the human behavior of annotating roads by tracing it at a specific orientation. We also develop a stacked multi-branch convolutional module to effectively utilize the mutual information between orientation learning and segmentation tasks. These contributions ensure that the model predicts topologically correct and connected road masks. We also propose Connectivity Refinement approach to further enhance the estimated road networks. The refinement model is pre-trained to connect and refine the corrupted ground-truth masks and later fine-tuned to enhance the predicted road masks. We demonstrate the advantages of our approach on two diverse road extraction datasets SpaceNet and DeepGlobe. Our approach improves over the state-of-the-art techniques by 9% and 7.5% in road topology metric on SpaceNet and DeepGlobe, respectively.

****

Deep Supervised Cross-Modal Retrieval

Liangli Zhen, Peng Hu, Xu Wang, Dezhong Peng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10394-10403

Cross-modal retrieval aims to enable flexible retrieval across different modalities. The core of cross-modal retrieval is how to measure the content similarity between different types of data. In this paper, we present a novel cross-modal retrieval method, called Deep Supervised Cross-modal Retrieval (DSCMR). It aims to find a common representation space, in which the samples from different modalities can be compared directly. Specifically, DSCMR minimises the discrimination loss in both the label space and the common representation space to supervise the model learning discriminative features. Furthermore, it simultaneously minimises the modality invariance loss and uses a weight sharing strategy to eliminate the cross-modal discrepancy of multimedia data in the common representation space to learn modality-invariant features. Comprehensive experimental results on four widely-used benchmark datasets demonstrate that the proposed method is effective in cross-modal learning and significantly outperforms the state-of-the-art cross-modal retrieval methods.

****

A Theoretically Sound Upper Bound on the Triplet Loss for Improving the Efficiency of Deep Distance Metric Learning

Thanh-Toan Do, Toan Tran, Ian Reid, Vijay Kumar, Tuan Hoang, Gustavo Carneiro; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10404-10413

We propose a method that substantially improves the efficiency of deep distance metric learning based on the optimization of the triplet loss function. One epoch of such training process based on a na"ive optimization of the triplet loss function has a run-time complexity O(N^3), where N is the number of training samples. Such optimization scales poorly, and the most common approach proposed to address this high complexity issue is based on sub-sampling the set of triplets needed for the training process. Another approach explored in the field relies on an ad-hoc linearization (in terms of N) of the triplet loss that introduces class centroids, which must be optimized using the whole training set for each mini-batch - this means that a na"ive implementation of this approach has run-time complexity O(N^2). This complexity issue is usually mitigated with poor, but computationally cheap, approximate centroid optimization methods. In this paper, we first propose a solid theory on the linearization of the triplet loss with the use of class centroids, where the main conclusion is that our new linear loss represents a tight upper-bound to the triplet loss. Furthermore, based on the theory above, we propose a training algorithm that no longer requires the centroid optimization step, which means that our approach is the first in the field with a guaranteed linear run-time complexity. We show that the training of deep distance metric learning methods using the proposed upper-bound is substantially faster than triplet-based methods, while producing competitive retrieval accuracy results on benchmark datasets (CUB-200-2011 and CAR196).

****
Data Representation and Learning With Graph Diffusion-Embedding Networks
Bo Jiang, Doudou Lin, Jin Tang, Bin Luo; Proceedings of the IEEE/CVF Conferen ce on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10414-10423
Recently, graph convolutional neural networks have been widely studied for graph -structured data representation and learning. In this paper, we present Graph Di ffusion-Embedding networks (GDENs), a new model for graph-structured data repres entation and learning. GDENs are motivated by our development of graph based fea ture diffusion. GDENs integrate both feature diffusion and graph node (low-dimen sional) embedding simultaneously into a unified network by employing a novel dif fusion-embedding architecture. GDENs have two main advantages. First, the equili brium representation of the diffusion-embedding operation in GDENs can be obtain ed via a simple closed-form solution, which thus guarantees the compactivity and efficiency of GDENs. Second, the proposed GDENs can be naturally extended to ad dress the data with multiple graph structures. Experiments on various semi-super vised learning tasks on several benchmark datasets demonstrate that the proposed GDENs significantly outperform traditional graph convolutional networks.
****
Video Relationship Reasoning Using Gated Spatio-Temporal Energy Graph
Yao-Hung Hubert Tsai, Santosh Divvala, Louis-Philippe Morency, Ruslan Salakhu tdinov, Ali Farhadi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10424-10433
Visual relationship reasoning is a crucial yet challenging task for understandin g rich interactions across visual concepts. For example, a relationship \ man, o pen, door\ involves a complex relation \ open\ between concrete entities \ man , door\ . While much of the existing work has studied this problem in the contex t of still images, understanding visual relationships in videos has received lim ited attention. Due to their temporal nature, videos enable us to model and reas on about a more comprehensive set of visual relationships, such as those requiri ng multiple (temporal) observations (e.g., \ man, lift up, box\ vs. \ man, put down, box\ ), as well as relationships that are often correlated through time (e .g., \ woman, pay, money\ followed by \ woman, buy, coffee\ ). In this paper, w e construct a Conditional Random Field on a fully-connected spatio-temporal grap h that exploits the statistical dependency between relational entities spatially and temporally. We introduce a novel gated energy function parametrization that learns adaptive relations conditioned on visual observations. Our model optimiz ation is computationally efficient, and its space computation complexity is sign ificantly amortized through our proposed parameterization. Experimental results on benchmark video datasets (ImageNet Video and Charades) demonstrate state-of-t he-art performance across three standard relationship reasoning tasks: Detection , Tagging, and Recognition.
****
Image-Question-Answer Synergistic Network for Visual Dialog
Dalu Guo, Chang Xu, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Com puter Vision and Pattern Recognition (CVPR), 2019, pp. 10434-10443
The image, question (combined with the history for de-referencing), and the corr esponding answer are three vital components of visual dialog. Classical visual d ialog systems integrate the image, question, and history to search for or genera te the best matched answer, and so, this approach significantly ignores the role of the answer. In this paper, we devise a novel image-question-answer synergist ic network to value the role of the answer for precise visual dialog. We extend the traditional one-stage solution to a two-stage solution. In the first stage, candidate answers are coarsely scored according to their relevance to the image and question pair. Afterward, in the second stage, answers with high probability of being correct are re-ranked by synergizing with image and question. On the V isual Dialog v1.0 dataset, the proposed synergistic network boosts the discrimin ative visual dialog model to achieve a new state-of-the-art of 57.88% normalized discounted cumulative gain. A generative visual dialog model equipped with the proposed technique also shows promising improvements.
****

Not All Frames Are Equal: Weakly-Supervised Video Grounding With Contextual Similarity and Visual Clustering Losses
Jing Shi, Jia Xu, Boqing Gong, Chenliang Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10444-10452
We invest the problem of weakly-supervised video grounding, where only video-level sentences are provided. This is a challenging task, and previous Multi-Instance Learning (MIL) based image grounding methods turn to fail in the video domain. Recent work attempts to decompose the video-level MIL into frame-level MIL by applying weighted sentence-frame ranking loss over frames, but it is not robust and does not exploit the rich temporal information in videos. In this work, we address these issues by extending frame-level MIL with a false positive frame-bag constraint and modeling the visual feature consistency in the video. In specific, we design a contextual similarity between semantic and visual features to deal with sparse objects association across frames. Furthermore, we leverage temporal coherence by strengthening the clustering effect of similar features in the visual space. We conduct an extensive evaluation on YouCookII and RoboWatch datasets, and demonstrate our method significantly outperforms prior state-of-the-art methods.
****
Inverse Cooking: Recipe Generation From Food Images
Amaia Salvador, Michal Drozdzal, Xavier Giro-i-Nieto, Adriana Romero; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10453-10462
People enjoy food photography because they appreciate food. Behind each meal there is a story described in a complex recipe and, unfortunately, by simply looking at a food image we do not have access to its preparation process. Therefore, in this paper we introduce an inverse cooking system that recreates cooking recipes given food images. Our system predicts ingredients as sets by means of a novel architecture, modeling their dependencies without imposing any order, and then generates cooking instructions by attending to both image and its inferred ingredients simultaneously. We extensively evaluate the whole system on the large-scale Recipe1M dataset and show that (1) we improve performance w.r.t. previous baselines for ingredient prediction; (2) we are able to obtain high quality recipes by leveraging both image and ingredients; (3) our system is able to produce more compelling recipes than retrieval-based approaches according to human judgment. We make code and models publicly available.
****
Adversarial Semantic Alignment for Improved Image Captions
Pierre Dognin, Igor Melnyk, Youssef Mroueh, Jerret Ross, Tom Sercu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10463-10471
In this paper, we study image captioning as a conditional GAN training, proposing both a context-aware LSTM captioner and co-attentive discriminator, which enforces semantic alignment between images and captions. We empirically focus on the viability of two training methods: Self-critical Sequence Training (SCST) and Gumbel Straight-Through (ST) and demonstrate that SCST shows more stable gradient behavior and improved results over Gumbel ST, even without accessing discriminator gradients directly. We also address the problem of automatic evaluation for captioning models and introduce a new semantic score, and show its correlation to human judgement. As an evaluation paradigm, we argue that an important criterion for a captioner is the ability to generalize to compositions of objects that do not usually co-occur together. To this end, we introduce a small captioned Out of Context (OOC) test set. The OOC set, combined with our semantic score, are the proposed new diagnosis tools for the captioning community. When evaluated on OOC and MS-COCO benchmarks, we show that SCST-based training has a strong performance in both semantic score and human evaluation, promising to be a valuable new approach for efficient discrete GAN training.
****
Answer Them All! Toward Universal Visual Question Answering Models
Robik Shrestha, Kushal Kafle, Christopher Kanan; Proceedings of the IEEE/CVF C

onference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10472-104
81
Visual Question Answering (VQA) research is split into two camps: the first focu
ses on VQA datasets that require natural image understanding and the second focu
ses on synthetic datasets that test reasoning. A good VQA algorithm should be ca
pable of both, but only a few VQA algorithms are tested in this manner. We compa
re five state-of-the-art VQA algorithms across eight VQA datasets covering both
domains. To make the comparison fair, all of the models are standardized as much
as possible, e.g., they use the same visual features, answer vocabularies, etc.
We find that methods do not generalize across the two domains. To address this
problem, we propose a new VQA algorithm that rivals or exceeds the state-of-the-
art for both domains.
****

Unsupervised Multi-Modal Neural Machine Translation
Yuanhang Su, Kai Fan, Nguyen Bach, C.-C. Jay Kuo, Fei Huang; Proceedings of
the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019,
pp. 10482-10491
Unsupervised neural machine translation (UNMT) has recently achieved remarkable
results [??] with only large monolingual corpora in each language. However, the
uncertainty of associating target with source sentences makes UNMT theoreticall
y an ill-posed problem. This work investigates the possibility of utilizing ima
ges for disambiguation to improve the performance of UNMT. Our assumption is in
tuitively based on the invariant property of image, i.e., the description of the
same visual content by different languages should be approximately similar. We
propose an unsupervised multi-modal machine translation (UMNMT) framework based
on the language translation cycle consistency loss conditional on the image, ta
rgeting to learn the bidirectional multi-modal translation simultaneously. Thro
ugh an alternate training between multi-modal and uni-modal, our inference model
can translate with or without the image. On the widely used Multi30K dataset,
the experimental results of our approach are significantly better than those of
the text-only UNMT on the 2016 test dataset.
****

Multi-Task Learning of Hierarchical Vision-Language Representation
Duy-Kien Nguyen, Takayuki Okatani; Proceedings of the IEEE/CVF Conference on Co
mputer Vision and Pattern Recognition (CVPR), 2019, pp. 10492-10501
It is still challenging to build an AI system that can perform tasks that involv
e vision and language at human level. So far, researchers have singled out indiv
idual tasks separately, for each of which they have designed networks and traine
d them on its dedicated datasets. Although this approach has seen a certain degr
ee of success, it comes with difficulties of understanding relations among diffe
rent tasks and transferring the knowledge learned for a task to others. We propo
se a multi-task learning approach that enables to learn vision-language represen
tation that is shared by many tasks from their diverse datasets. The representat
ion is hierarchical, and prediction for each task is computed from the represent
ation at its corresponding level of the hierarchy. We show through experiments
that our method consistently outperforms previous single-task-learning methods o
n image caption retrieval, visual question answering, and visual grounding. We a
lso analyze the learned hierarchical representation by visualizing attention map
s generated in our network.
****

Cross-Modal Self-Attention Network for Referring Image Segmentation
Linwei Ye, Mrigank Rochan, Zhi Liu, Yang Wang; Proceedings of the IEEE/CVF Co
nference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10502-1051
1
We consider the problem of referring image segmentation. Given an input image an
d a natural language expression, the goal is to segment the object referred by t
he language expression in the image. Existing works in this area treat the langu
age expression and the input image separately in their representations. They do
not sufficiently capture long-range correlations between these two modalities. I
n this paper, we propose a cross-modal self-attention (CMSA) module that effecti

vely captures the long-range dependencies between linguistic and visual features . Our model can adaptively focus on informative words in the referring expressio n and important regions in the input image. In addition, we propose a gated mult i-level fusion module to selectively integrate self-attentive cross-modal featur es corresponding to different levels in the image. This module controls the info rmation flow of features at different levels. We validate the proposed approach on four evaluation datasets. Our proposed approach consistently outperforms exis ting state-of-the-art methods.
****

DuDoNet: Dual Domain Network for CT Metal Artifact Reduction
Wei-An Lin, Haofu Liao, Cheng Peng, Xiaohang Sun, Jingdan Zhang, Jiebo Luo, Rama Chellappa, Shaohua Kevin Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10512-10521
Computed tomography (CT) is an imaging modality widely used for medical diagnosi s and treatment. CT images are often corrupted by undesirable artifacts when met allic implants are carried by patients, which creates the problem of metal artif act reduction (MAR). Existing methods for reducing the artifacts due to metallic implants are inadequate for two main reasons. First, metal artifacts are struct ured and non-local so that simple image domain enhancement approaches would not suffice. Second, the MAR approaches which attempt to reduce metal artifacts in t he X-ray projection (sinogram) domain inevitably lead to severe secondary artifa ct due to sinogram inconsistency. To overcome these difficulties, we propose an end-to-end trainable Dual Domain Network (DuDoNet) to simultaneously restore sin ogram consistency and enhance CT images. The linkage between the sigogram and im age domains is a novel Radon inversion layer that allows the gradients to back-p ropagate from the image domain to the sinogram domain during training. Extensive experiments show that our method achieves significant improvements over other s ingle domain MAR approaches. To the best of our knowledge, it is the first end-t o-end dual-domain network for MAR.
****

Fast Spatio-Temporal Residual Network for Video Super-Resolution
Sheng Li, Fengxiang He, Bo Du, Lefei Zhang, Yonghao Xu, Dacheng Tao; Procee dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP R), 2019, pp. 10522-10531
Recently, deep learning based video super-resolution (SR) methods have achieved promising performance. To simultaneously exploit the spatial and temporal inform ation of videos, employing 3-dimensional (3D) convolutions is a natural approach . However, straight utilizing 3D convolutions may lead to an excessively high co mputational complexity which restricts the depth of video SR models and thus und ermine the performance. In this paper, we present a novel fast spatio-temporal r esidual network (FSTRN) to adopt 3D convolutions for the video SR task in order to enhance the performance while maintaining a low computational load. Specifica lly, we propose a fast spatio-temporal residual block (FRB) that divide each 3D filter to the product of two 3D filters, which have considerably lower dimension s. Furthermore, we design a cross-space residual learning that directly links th e low-resolution space and the high-resolution space, which can greatly relieve the computational burden on the feature fusion and up-scaling parts. Extensive e valuations and comparisons on benchmark datasets validate the strengths of the p roposed approach and demonstrate that the proposed network significantly outperf orms the current state-of-the-art methods.
****

Complete the Look: Scene-Based Complementary Product Recommendation
Wang-Cheng Kang, Eric Kim, Jure Leskovec, Charles Rosenberg, Julian McAuley; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti on (CVPR), 2019, pp. 10532-10541
Modeling fashion compatibility is challenging due to its complexity and subjecti vity. Existing work focuses on predicting compatibility between product images (e.g. an image containing a t-shirt and an image containing a pair of jeans). How ever, these approaches ignore real-world 'scene' images (e.g. selfies); such ima ges are hard to deal with due to their complexity, clutter, variations in lighti

ng and pose (etc.) but on the other hand could potentially provide key context (e.g. the user's body type, or the season) for making more accurate recommendatio ns. In this work, we propose a new task called 'Complete the Look', which seeks to recommend visually compatible products based on scene images. We design an ap proach to extract training data for this task, and propose a novel way to learn the scene-product compatibility from fashion or interior design images. Our appr oach measures compatibility both globally and locally via CNNs and attention mec hanisms. Extensive experiments show that our method achieves significant perform ance gains over alternative systems. Human evaluation and qualitative analysis a re also conducted to further understand model behavior. We hope this work could lead to useful applications which link large corpora of real-world scenes with s hoppable products.
****

Selective Sensor Fusion for Neural Visual-Inertial Odometry
Changhao Chen, Stefano Rosa, Yishu Miao, Chris Xiaoxuan Lu, Wei Wu, Andrew Markham, Niki Trigoni; Proceedings of the IEEE/CVF Conference on Computer Visio n and Pattern Recognition (CVPR), 2019, pp. 10542-10551
Deep learning approaches for Visual-Inertial Odometry (VIO) have proven successf ul, but they rarely focus on incorporating robust fusion strategies for dealing with imperfect input sensory data. We propose a novel end-to-end selective senso r fusion framework for monocular VIO, which fuses monocular images and inertial measurements in order to estimate the trajectory whilst improving robustness to real-life issues, such as missing and corrupted data or bad sensor synchronizati on. In particular, we propose two fusion modalities based on different masking s trategies: deterministic soft fusion and stochastic hard fusion, and we compare with previously proposed direct fusion baselines. During testing, the network is able to selectively process the features of the available sensor modalities and produce a trajectory at scale. We present a thorough investigation on the perf ormances on three public autonomous driving, Micro Aerial Vehicle (MAV) and hand -held VIO datasets. The results demonstrate the effectiveness of the fusion stra tegies, which offer better performances compared to direct fusion, particularly in presence of corrupted data. In addition, we study the interpretability of the fusion networks by visualising the masking layers in different scenarios and wi th varying data corruption, revealing interesting correlations between the fusio n networks and imperfect sensory input data.
****

Look More Than Once: An Accurate Detector for Text of Arbitrary Shapes
Chengquan Zhang, Borong Liang, Zuming Huang, Mengyi En, Junyu Han, Errui Di ng, Xinghao Ding; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10552-10561
Previous scene text detection methods have progressed substantially over the pas t years. However, limited by the receptive field of CNNs and the simple represen tations like rectangle bounding box or quadrangle adopted to describe text, prev ious methods may fall short when dealing with more challenging text instances, s uch as extremely long text and arbitrarily shaped text. To address these two pro blems, we present a novel text detector namely LOMO, which localizes the text pr ogressively for multiple times (or in other word, LOok More than Once). LOMO con sists of a direct regressor (DR), an iterative refinement module (IRM) and a sha pe expression module (SEM). At first, text proposals in the form of quadrangle a re generated by DR branch. Next, IRM progressively perceives the entire long tex t by iterative refinement based on the extracted feature blocks of preliminary p roposals. Finally, a SEM is introduced to reconstruct more precise representatio n of irregular text by considering the geometry properties of text instance, inc luding text region, text center line and border offsets. The state-of-the-art re sults on several public benchmarks including ICDAR2017-RCTW, SCUT-CTW1500, Total -Text, ICDAR2015 and ICDAR17-MLT confirm the striking robustness and effectivene ss of LOMO.
****

Learning Binary Code for Personalized Fashion Recommendation
Zhi Lu, Yang Hu, Yunchao Jiang, Yan Chen, Bing Zeng; Proceedings of the IEEE

/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10562-10570
With the rapid growth of fashion-focused social networks and online shopping, intelligent fashion recommendation is now in great needs. Recommending fashion outfits, each of which is composed of multiple interacted clothing and accessories, is relatively new to the field. The problem becomes even more interesting and challenging when considering users' personalized fashion style. Another challenge in a large-scale fashion outfit recommendation system is the efficiency issue of item/outfit search and storage. In this paper, we propose to learn binary code for efficient personalized fashion outfits recommendation. Our system consists of three components, a feature network for content extraction, a set of type-dependent hashing modules to learn binary codes, and a matching block that conducts pairwise matching. The whole framework is trained in an end-to-end manner. We collect outfit data together with user label information from a fashion-focused social website for the personalized recommendation task. Extensive experiments on our datasets show that the proposed framework outperforms the state-of-the-art methods significantly even with a simple backbone.
****

Attention Based Glaucoma Detection: A Large-Scale Database and CNN Model
Liu Li, Mai Xu, Xiaofei Wang, Lai Jiang, Hanruo Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10571-10580
Recently, the attention mechanism has been successfully applied in convolutional neural networks (CNNs), significantly boosting the performance of many computer vision tasks. Unfortunately, few medical image recognition approaches incorporate the attention mechanism in the CNNs. In particular, there exists high redundancy in fundus images for glaucoma detection, such that the attention mechanism has potential in improving the performance of CNN-based glaucoma detection. This paper proposes an attention-based CNN for glaucoma detection (AG-CNN). Specifically, we first establish a large-scale attention based glaucoma (LAG) database, which includes 5,824 fundus images labeled with either positive glaucoma (2,392) or negative glaucoma (3,432). The attention maps of the ophthalmologists are also collected in LAG database through a simulated eye-tracking experiment. Then, a new structure of AG-CNN is designed, including an attention prediction subnet, a pathological area localization subnet and a glaucoma classification subnet. Different from other attention-based CNN methods, the features are also visualized as the localized pathological area, which can advance the performance of glaucoma detection. Finally, the experiment results show that the proposed AG-CNN approach significantly advances state-of-the-art glaucoma detection.
****

Privacy Protection in Street-View Panoramas Using Depth and Multi-View Imagery
Ries Uittenbogaard, Clint Sebastian, Julien Vijverberg, Bas Boom, Dariu M. Gavrila, Peter H.N. de With; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10581-10590
The current paradigm in privacy protection in street-view images is to detect and blur sensitive information. In this paper, we propose a framework that is an alternative to blurring, which automatically removes and inpaints moving objects (e.g. pedestrians, vehicles) in street-view imagery. We propose a novel moving object segmentation algorithm exploiting consistencies in depth across multiple street-view images that are later combined with the results of a segmentation network. The detected moving objects are removed and inpainted with information from other views, to obtain a realistic output image such that the moving object is not visible anymore. We evaluate our results on a dataset of 1000 images to obtain a peak noise-to-signal ratio (PSNR) and L 1 loss of 27.2 dB and 2.5%, respectively. To assess overall quality, we also report the results of a survey conducted on 35 professionals, asked to visually inspect the images whether object removal and inpainting had taken place. The inpainting dataset will be made publicly available for scientific benchmarking purposes at https://research.cyclomedia.com/.
****

Grounding Human-To-Vehicle Advice for Self-Driving Vehicles
Jinkyu Kim, Teruhisa Misu, Yi-Ting Chen, Ashish Tawari, John Canny; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10591-10599
Recent success suggests that deep neural control networks are likely to be a key component of self-driving vehicles. These networks are trained on large datasets to imitate human actions, but they lack semantic understanding of image contents. This makes them brittle and potentially unsafe in situations that do not match training data. Here, we propose to address this issue by augmenting training data with natural language advice from a human. Advice includes guidance about what to do and where to attend. We present the first step toward advice giving, where we train an end-to-end vehicle controller that accepts advice. The controller adapts the way it attends to the scene (visual attention) and the control (steering and speed). Attention mechanisms tie controller behavior to salient objects in the advice. We evaluate our model on a novel advisable driving dataset with manually annotated human-to-vehicle advice called Honda Research Institute-Advice Dataset (HAD). We show that taking advice improves the performance of the end-to-end network, while the network cues on a variety of visual features that are provided by advice. The dataset is available at https://usa.honda-ri.com/HAD.
****
Multi-Step Prediction of Occupancy Grid Maps With Recurrent Neural Networks
Nima Mohajerin, Mohsen Rohani; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10600-10608
We investigate the multi-step prediction of the drivable space, represented by Occupancy Grid Maps (OGMs), for autonomous vehicles. Our motivation is that accurate multi-step prediction of the drivable space can efficiently improve path planning and navigation resulting in safe, comfortable and optimum paths in autonomous driving. We train a variety of Recurrent Neural Network (RNN) based architectures on the OGM sequences from the KITTI dataset. The results demonstrate significant improvement of the prediction accuracy using our proposed difference learning method, incorporating motion related features, over the state of the art. We remove the egomotion from the OGM sequences by transforming them into a common frame. Although in the transformed sequences the KITTI dataset is heavily biased toward static objects, by learning the difference between consecutive OGMs, our proposed method provides accurate prediction over both the static and moving objects.
****
Connecting Touch and Vision via Cross-Modal Prediction
Yunzhu Li, Jun-Yan Zhu, Russ Tedrake, Antonio Torralba; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10609-10618
Humans perceive the world using multi-modal sensory inputs such as vision, audition, and touch. In this work, we investigate the cross-modal connection between vision and touch. The main challenge in this cross-domain modeling task lies in the significant scale discrepancy between the two: while our eyes perceive an entire visual scene at once, humans can only feel a small region of an object at any given moment. To connect vision and touch, we introduce new tasks of synthesizing plausible tactile signals from visual inputs as well as imagining how we interact with objects given tactile data as input. To accomplish our goals, we first equip robots with both visual and tactile sensors and collect a large-scale dataset of corresponding vision and tactile image sequences. To close the scale gap, we present a new conditional adversarial model that incorporates the scale and location information of the touch. Human perceptual studies demonstrate that our model can produce realistic visual images from tactile data and vice versa. Finally, we present both qualitative and quantitative experimental results regarding different system designs, as well as visualizing the learned representations of our model.
****
X2CT-GAN: Reconstructing CT From Biplanar X-Rays With Generative Adversarial Networks

Xingde Ying, Heng Guo, Kai Ma, Jian Wu, Zhengxin Weng, Yefeng Zheng; Procee dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP R), 2019, pp. 10619-10628

Computed tomography (CT) can provide a 3D view of the patient's internal organs, facilitating disease diagnosis, but it incurs more radiation dose to a patient and a CT scanner is much more cost prohibitive than an X-ray machine too. Tradit ional CT reconstruction methods require hundreds of X-ray projections through a full rotational scan of the body, which cannot be performed on a typical X-ray m achine. In this work, we propose to reconstruct CT from two orthogonal X-rays us ing the generative adversarial network (GAN) framework. A specially designed gen erator network is exploited to increase data dimension from 2D (X-rays) to 3D (C T), which is not addressed in previous research of GAN. A novel feature fusion m ethod is proposed to combine information from two X-rays. The mean squared error (MSE) loss and adversarial loss are combined to train the generator, resulting in a high-quality CT volume both visually and quantitatively. Extensive experime nts on a publicly available chest CT dataset demonstrate the effectiveness of th e proposed method. It could be a nice enhancement of a low-cost X-ray machine to provide physicians a CT-like 3D volume in several niche applications.

****

Practical Full Resolution Learned Lossless Image Compression

Fabian Mentzer, Eirikur Agustsson, Michael Tschannen, Radu Timofte, Luc Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Reco gnition (CVPR), 2019, pp. 10629-10638

We propose the first practical learned lossless image compression system, L3C, a nd show that it outperforms the popular engineered codecs, PNG, WebP and JPEG 20 00. At the core of our method is a fully parallelizable hierarchical probabilist ic model for adaptive entropy coding which is optimized end-to-end for the compr ession task. In contrast to recent autoregressive discrete probabilistic models such as PixelCNN, our method i) models the image distribution jointly with learn ed auxiliary representations instead of exclusively modeling the image distribut ion in RGB space, and ii) only requires three forward-passes to predict all pixe l probabilities instead of one for each pixel. As a result, L3C obtains over two orders of magnitude speedups when sampling compared to the fastest PixelCNN var iant (Multiscale-PixelCNN). Furthermore, we find that learning the auxiliary rep resentation is crucial and outperforms predefined auxiliary representations such as an RGB pyramid significantly.

****

Image-To-Image Translation via Group-Wise Deep Whitening-And-Coloring Transforma tion

Wonwoong Cho, Sungha Choi, David Keetae Park, Inkyu Shin, Jaegul Choo; Proce edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV PR), 2019, pp. 10639-10647

Recently, unsupervised exemplar-based image-to-image translation, conditioned on a given exemplar without the paired data, has accomplished substantial advancem ents. In order to transfer the information from an exemplar to an input image, e xisting methods often use a normalization technique, e.g., adaptive instance nor malization, that controls the channel-wise statistics of an input activation map at a particular layer, such as the mean and the variance. Meanwhile, style tran sfer approaches similar task to image translation by nature, demonstrated superi or performance by using the higher-order statistics such as covariance among cha nnels in representing a style. In detail, it works via whitening (given a zero-m ean input feature, transforming its covariance matrix into the identity). follow ed by coloring (changing the covariance matrix of the whitened feature to those of the style feature). However, applying this approach in image translation is c omputationally intensive and error-prone due to the expensive time complexity an d its non-trivial backpropagation. In response, this paper proposes an end-to-en d approach tailored for image translation that efficiently approximates this tra nsformation with our novel regularization methods. We further extend our approac h to a group-wise form for memory and time efficiency as well as image quality. Extensive qualitative and quantitative experiments demonstrate that our proposed

method is fast, both in training and inference, and highly effective in reflecting the style of an exemplar.
****

Max-Sliced Wasserstein Distance and Its Use for GANs
Ishan Deshpande, Yuan-Ting Hu, Ruoyu Sun, Ayis Pyrros, Nasir Siddiqui, Sanmi Koyejo, Zhizhen Zhao, David Forsyth, Alexander G. Schwing; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10648-10656
Generative adversarial nets (GANs) and variational auto-encoders have significantly improved our distribution modeling capabilities, showing promise for dataset augmentation, image-to-image translation and feature learning. However, to model high-dimensional distributions, sequential training and stacked architectures are common, increasing the number of tunable hyper-parameters as well as the training time. Nonetheless, the sample complexity of the distance metrics remains one of the factors affecting GAN training. We first show that the recently proposed sliced Wasserstein distance has compelling sample complexity properties when compared to the Wasserstein distance. To further improve the sliced Wasserstein distance we then analyze its `projection complexity' and develop the max-sliced Wasserstein distance which enjoys compelling sample complexity while reducing projection complexity, albeit necessitating a max estimation. We finally illustrate that the proposed distance trains GANs on high-dimensional images up to a resolution of 256x256 easily.
****

Meta-Learning With Differentiable Convex Optimization
Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, Stefano Soatto; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10657-10665
Many meta-learning approaches for few-shot learning rely on simple base learners such as nearest-neighbor classifiers. However, even in the few-shot regime, discriminatively trained linear predictors can offer better generalization. We propose to use these predictors as base learners to learn representations for few-shot learning and show they offer better tradeoffs between feature size and performance across a range of few-shot recognition benchmarks. Our objective is to learn feature embeddings that generalize well under a linear classification rule for novel categories. To efficiently solve the objective, we exploit two properties of linear classifiers: implicit differentiation of the optimality conditions of the convex problem and the dual formulation of the optimization problem. This allows us to use high-dimensional embeddings with improved generalization at a modest increase in computational overhead. Our approach, named MetaOptNet, achieves state-of-the-art performance on miniImageNet, tieredImageNet, CIFAR-FS, and FC100 few-shot learning benchmarks.
****

RePr: Improved Training of Convolutional Filters
Aaditya Prakash, James Storer, Dinei Florencio, Cha Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10666-10675
A well-trained Convolutional Neural Network can easily be pruned without significant loss of performance. This is because of unnecessary overlap in the features captured by the network's filters. Innovations in network architecture such as skip/dense connections and inception units have mitigated this problem to some extent, but these improvements come with increased computation and memory requirements at run-time. We attempt to address this problem from another angle - not by changing the network structure but by altering the training method. We show that by temporarily pruning and then restoring a subset of the model's filters, and repeating this process cyclically, overlap in the learned features is reduced, producing improved generalization. We show that the existing model-pruning criteria are not optimal for selecting filters to prune in this context, and introduce inter-filter orthogonality as the ranking criteria to determine under-expressive filters. Our method is applicable both to vanilla convolutional networks and more complex modern architectures, and improves the performance across a variet

y of tasks, especially when applied to smaller networks.
****

Tangent-Normal Adversarial Regularization for Semi-Supervised Learning
Bing Yu, Jingfeng Wu, Jinwen Ma, Zhanxing Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10676-10684
Compared with standard supervised learning, the key difficulty in semi-supervised learning is how to make full use of the unlabeled data. A recently proposed method, virtual adversarial training (VAT), smartly performs adversarial training without label information to impose a local smoothness on the classifier, which is especially beneficial to semi-supervised learning. In this work, we propose tangent-normal adversarial regularization (TNAR) as an extension of VAT by taking the data manifold into consideration. The proposed TNAR is composed by two complementary parts, the tangent adversarial regularization (TAR) and the normal adversarial regularization (NAR). In TAR, VAT is applied along the tangent space of the data manifold, aiming to enforce local invariance of the classifier on the manifold, while in NAR, VAT is performed on the normal space orthogonal to the tangent space, intending to impose robustness on the classifier against the noise causing the observed data deviating from the underlying data manifold. Demonstrated by experiments on both artificial and practical datasets, our proposed TAR and NAR complement with each other, and jointly outperforms other state-of-the-art methods for semi-supervised learning.
****

Auto-Encoding Scene Graphs for Image Captioning
Xu Yang, Kaihua Tang, Hanwang Zhang, Jianfei Cai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10685-10694
We propose Scene Graph Auto-Encoder (SGAE) that incorporates the language inductive bias into the encoder-decoder image captioning framework for more human-like captions. Intuitively, we humans use the inductive bias to compose collocations and contextual inference in discourse. For example, when we see the relation "person on bike", it is natural to replace "on" with "ride" and infer "person riding bike on a road" even the "road" is not evident. Therefore, exploiting such bias as a language prior is expected to help the conventional encoder-decoder models less likely to overfit to the dataset bias and focus on reasoning. Specifically, we use the scene graph --- a directed graph (G) where an object node is connected by adjective nodes and relationship nodes --- to represent the complex structural layout of both image (I) and sentence (S). In the textual domain, we use SGAE to learn a dictionary (D) that helps to reconstruct sentences in the S -> G -> D -> S pipeline, where D encodes the desired language prior; in the vision-language domain, we use the shared D to guide the encoder-decoder in the I -> G -> D -> S pipeline. Thanks to the scene graph representation and shared dictionary, the inductive bias is transferred across domains in principle. We validate the effectiveness of SGAE on the challenging MS-COCO image captioning benchmark, e.g., our SGAE-based single-model achieves a new state-of-the-art 127.8 CIDEr-D on the Karpathy split, and a competitive 125.5 CIDEr-D (c40) on the official server even compared to other ensemble models. Code has been made available at: https://github.com/yangxuntu/SGAE.
****

Fast, Diverse and Accurate Image Captioning Guided by Part-Of-Speech
Aditya Deshpande, Jyoti Aneja, Liwei Wang, Alexander G. Schwing, David Forsyth; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10695-10704
Image captioning is an ambiguous problem, with many suitable captions for an image. To address ambiguity, beam search is the de facto method for sampling multiple captions. However, beam search is computationally expensive and known to produce generic captions. To address this concern, some variational auto-encoder (VAE) and generative adversarial net (GAN) based methods have been proposed. Though diverse, GAN and VAE are less accurate. In this paper, we first predict a meaningful summary of the image, then generate the caption based on that summa

ry. We use part-of-speech as summaries, since our summary should drive caption g
eneration. We achieve the trifecta: (1) High accuracy for the diverse captions
as evaluated by standard captioning metrics and user studies; (2) Faster comput
ation of diverse captions compared to beam search and diverse beam search; and
(3) High diversity as evaluated by counting novel sentences, distinct n-grams
and mutual overlap (i.e., mBleu-4) scores.
****

Attention Branch Network: Learning of Attention Mechanism for Visual Explanation
Hiroshi Fukui, Tsubasa Hirakawa, Takayoshi Yamashita, Hironobu Fujiyoshi; Pro
ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (
CVPR), 2019, pp. 10705-10714
Visual explanation enables humans to understand the decision making of deep conv
olutional neural network (CNN), but it is insufficient to contribute to improvin
g CNN performance. In this paper, we focus on the attention map for visual expla
nation, which represents a high response value as the attention location in imag
e recognition. This attention region significantly improves the performance of C
NN by introducing an attention mechanism that focuses on a specific region in an
image. In this work, we propose Attention Branch Network (ABN), which extends a
response-based visual explanation model by introducing a branch structure with
an attention mechanism. ABN can be applicable to several image recognition tasks
by introducing a branch for the attention mechanism and is trainable for visual
explanation and image recognition in an end-to-end manner. We evaluate ABN on
several image recognition tasks such as image classification, fine-grained recog
nition, and multiple facial attribute recognition. Experimental results indicate
that ABN outperforms the baseline models on these image recognition tasks while
generating an attention map for visual explanation. Our code is available.
****

Cascaded Projection: End-To-End Network Compression and Acceleration
Breton Minnehan, Andreas Savakis; Proceedings of the IEEE/CVF Conference on Com
puter Vision and Pattern Recognition (CVPR), 2019, pp. 10715-10724
We propose a data-driven approach for deep convolutional neural network compress
ion that achieves high accuracy with high throughput and low memory requirements
. Current network compression methods either find a low-rank factorization of th
e features that requires more memory or select only a subset of features by prun
ing entire filter channels. We propose the Cascaded Projection (CaP) compression
method that projects the output and input filter channels of successive layers
to a unified low dimensional space based on a low-rank projection. We optimize t
he projection to minimize classification loss and the difference between the nex
t layer's features in the compressed and uncompressed networks. To solve this no
n-convex optimization problem we propose a new optimization method of a proxy ma
trix using backpropagation and Stochastic Gradient Descent (SGD) with geometric
constraints. Our cascaded projection approach leads to improvements in all criti
cal areas of network compression: high accuracy, low memory consumption, low par
ameter count and high processing speed. The proposed CaP method demonstrates sta
te of the art results compressing VGG16 and ResNet networks with over 4X reducti
on in the number of computations and excellent performance in top-5 accuracy on
the ImageNet dataset before and after fine-tuning.
****

DeepCaps: Going Deeper With Capsule Networks
Jathushan Rajasegaran, Vinoj Jayasundara, Sandaru Jayasekara, Hirunima Jayase
kara, Suranga Seneviratne, Ranga Rodrigo; Proceedings of the IEEE/CVF Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10725-10733
Capsule Network is a promising concept in deep learning, yet its true potential
is not fully realized thus far, providing sub-par performance on several key ben
chmark datasets with complex data. Drawing intuition from the success achieved b
y Convolutional Neural Networks (CNNs) by going deeper, we introduce DeepCaps, a
deep capsule network architecture which uses a novel 3D convolution based dynam
ic routing algorithm. With DeepCaps, we surpass the state-of-the-art capsule dom
ain networks results on CIFAR10, SVHN and Fashion MNIST, while achieving a 68% r
eduction in the number of parameters. Further, we propose a class independent de

coder network, which strengthens the use of reconstruction loss as a regularizat
ion term. This leads to an interesting property of the decoder, which allows us
to identify and control the physical attributes of the images represented by the
 instantiation parameters.
****
FBNet: Hardware-Aware Efficient ConvNet Design via Differentiable Neural Archite
cture Search
Bichen Wu, Xiaoliang Dai, Peizhao Zhang, Yanghan Wang, Fei Sun, Yiming Wu,
 Yuandong Tian, Peter Vajda, Yangqing Jia, Kurt Keutzer; Proceedings of the I
EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp.
10734-10742
Designing accurate and efficient ConvNets for mobile devices is challenging beca
use the design space is combinatorially large. Due to this, previous neural arch
itecture search (NAS) methods are computationally expensive. ConvNet architectur
e optimality depends on factors such as input resolution and target devices. How
ever, existing approaches are too resource demanding for case-by-case redesigns.
 Also, previous work focuses primarily on reducing FLOPs, but FLOP count does no
t always reflect actual latency. To address these, we propose a differentiable n
eural architecture search (DNAS) framework that uses gradient-based methods to o
ptimize ConvNet architectures, avoiding enumerating and training individual arch
itectures separately as in previous methods. FBNets (Facebook-Berkeley-Nets), a
family of models discovered by DNAS surpass state-of-the-art models both designe
d manually and generated automatically. FBNet-B achieves 74.1% top-1 accuracy on
 ImageNet with 295M FLOPs and 23.1 ms latency on a Samsung S8 phone, 2.4x smalle
r and 1.5x faster than MobileNetV2-1.3 with similar accuracy. Despite higher acc
uracy and lower latency than MnasNet, we estimate FBNet-B's search cost is 420x
smaller than MnasNet's, at only 216 GPU-hours. Searched for different resolution
s and channel sizes, FBNets achieve 1.5% to 6.4% higher accuracy than MobileNetV
2. The smallest FBNet achieves 50.2% accuracy and 2.9 ms latency (345 frames per
 second) on a Samsung S8. Over a Samsung-optimized FBNet, the iPhone-X-optimized
 model achieves a 1.4x speedup on an iPhone X. FBNet models are open-sourced at
https://github. com/facebookresearch/mobile-vision.
****
APDrawingGAN: Generating Artistic Portrait Drawings From Face Photos With Hierar
chical GANs
Ran Yi, Yong-Jin Liu, Yu-Kun Lai, Paul L. Rosin; Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10743-10
752
Significant progress has been made with image stylization using deep learning, e
specially with generative adversarial networks (GANs). However, existing methods
 fail to produce high quality artistic portrait drawings. Such drawings have a h
ighly abstract style, containing a sparse set of continuous graphical elements s
uch as lines, and so small artifacts are much more exposed than for painting sty
les. Moreover, artists tend to use different strategies to draw different facial
 features and the lines drawn are only loosely related to obvious image features
. To address these challenges, we propose APDrawingGAN, a novel GAN based archit
ecture that builds upon hierarchical generators and discriminators combining bot
h a global network (for images as a whole) and local networks (for individual fa
cial regions). This allows dedicated drawing strategies to be learned for differ
ent facial features. Since artists' drawings may not have lines perfectly aligne
d with image features, we develop a novel loss to measure similarity between gen
erated and artists' drawings based on distance transforms, leading to improved s
trokes in portrait drawing. To train APDrawingGAN, we construct an artistic draw
ing dataset containing high-resolution portrait photos and corresponding profess
ional artistic drawings. Extensive experiments, including a user study, show tha
t APDrawingGAN produces significantly better artistic drawings than state-of-the
-art methods.
****
Constrained Generative Adversarial Networks for Interactive Image Generation
Eric Heim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern

Recognition (CVPR), 2019, pp. 10753-10761
Generative Adversarial Networks (GANs) have received a great deal of attention due in part to recent success in generating original, high-quality samples from visual domains. However, most current methods only allow for users to guide this image generation process through limited interactions. In this work we develop a novel GAN framework that allows humans to be "in-the-loop" of the image generation process. Our technique iteratively accepts relative constraints of the form "Generate an image more like image A than image B". After each constraint is given, the user is presented with new outputs from the GAN, informing the next round of feedback. This feedback is used to constrain the output of the GAN with respect to an underlying semantic space that can be designed to model a variety of different notions of similarity (e.g. classes, attributes, object relationships, color, etc.). In our experiments, we show that our GAN framework is able to generate images that are of comparable quality to equivalent unsupervised GANs while satisfying a large number of the constraints provided by users, effectively changing a GAN into one that allows users interactive control over image generation without sacrificing image quality.
****

WarpGAN: Automatic Caricature Generation
Yichun Shi, Debayan Deb, Anil K. Jain; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10762-10771
We propose, WarpGAN, a fully automatic network that can generate caricatures given an input face photo. Besides transferring rich texture styles, WarpGAN learns to automatically predict a set of control points that can warp the photo into a caricature, while preserving identity. We introduce an identity-preserving adversarial loss that aids the discriminator to distinguish between different subjects. Moreover, WarpGAN allows customization of the generated caricatures by controlling the exaggeration extent and the visual styles. Experimental results on a public domain dataset, WebCaricature, show that WarpGAN is capable of generating caricatures that not only preserve the identities but also outputs a diverse set of caricatures for each input photo. Five caricature experts suggest that caricatures generated by WarpGAN are visually similar to hand-drawn ones and only prominent facial features are exaggerated.
****

Explainability Methods for Graph Convolutional Neural Networks
Phillip E. Pope, Soheil Kolouri, Mohammad Rostami, Charles E. Martin, Heiko Hoffmann; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10772-10781
With the growing use of graph convolutional neural networks (GCNNs) comes the need for explainability. In this paper, we introduce explainability methods for GCNNs. We develop the graph analogues of three prominent explainability methods for convolutional neural networks: contrastive gradient-based (CG) saliency maps, Class Activation Mapping (CAM), and Excitation Back-Propagation (EB) and their variants, gradient-weighted CAM (Grad-CAM) and contrastive EB (c-EB). We show a proof-of-concept of these methods on classification problems in two application domains: visual scene graphs and molecular graphs. To compare the methods, we identify three desirable properties of explanations: (1) their importance to classification, as measured by the impact of occlusions, (2) their contrastivity with respect to different classes, and (3) their sparseness on a graph. We call the corresponding quantitative metrics fidelity, contrastivity, and sparsity and evaluate them for each method. Lastly, we analyze the salient subgraphs obtained from explanations and report frequently occurring patterns.
****

A Generative Adversarial Density Estimator
M. Ehsan Abbasnejad, Qinfeng Shi, Anton van den Hengel, Lingqiao Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10782-10791
Density estimation is a challenging unsupervised learning problem. Current maximum likelihood approaches for density estimation are either restrictive or incapable of producing high-quality samples. On the other hand, likelihood-free models

such as generative adversarial networks, produce sharp samples without a density model. The lack of a density estimate limits the applications to which the sampled data can be put, however. We propose a Generative Adversarial Density Estimator, a density estimation approach that bridges the gap between the two. Allowing for a prior on the parameters of the model, we extend our density estimator to a Bayesian model where we can leverage the predictive variance to measure our confidence in the likelihood. Our experiments on challenging applications such as visual dialog where the density and the confidence in predictions are crucial shows the effectiveness of our approach.
****

SoDeep: A Sorting Deep Net to Learn Ranking Loss Surrogates
Martin Engilberge, Louis Chevallier, Patrick Perez, Matthieu Cord; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10792-10801
Several tasks in machine learning are evaluated using non-differentiable metrics such as mean average precision or Spearman correlation. However, their non-differentiability prevents from using them as objective functions in a learning framework. Surrogate and relaxation methods exist but tend to be specific to a given metric. In the present work, we introduce a new method to learn approximations of such non-differentiable objective functions. Our approach is based on a deep architecture that approximates the sorting of arbitrary sets of scores. It is trained virtually for free using synthetic data. This sorting deep (SoDeep) net can then be combined in a plug-and-play manner with existing deep architectures. We demonstrate the interest of our approach in three different tasks that require ranking: Cross-modal text-image retrieval, multi-label image classification and visual memorability ranking. Our approach yields very competitive results on these three tasks, which validates the merit and the flexibility of SoDeep as a proxy for sorting operation in ranking-based losses.
****

High-Quality Face Capture Using Anatomical Muscles
Michael Bao, Matthew Cong, Stephane Grabli, Ronald Fedkiw; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10802-10811
Muscle-based systems have the potential to provide both anatomical accuracy and semantic interpretability as compared to blendshape models; however, a lack of expressivity and differentiability has limited their impact. Thus, we propose modifying a recently developed rather expressive muscle-based system in order to make it fully-differentiable; in fact, our proposed modifications allow this physically robust and anatomically accurate muscle model to conveniently be driven by an underlying blendshape basis. Our formulation is intuitive, natural, as well as monolithically and fully coupled such that one can differentiate the model from end to end, which makes it viable for both optimization and learning-based approaches for a variety of applications. We illustrate this with a number of examples including both shape matching of three-dimensional geometry as as well as the automatic determination of a three-dimensional facial pose from a single two-dimensional RGB image without using markers or depth information.
****

FML: Face Model Learning From Videos
Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Perez, Michael Zollhofer, Christian Theobalt; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10812-10822
Monocular image-based 3D reconstruction of faces is a long-standing problem in computer vision. Since image data is a 2D projection of a 3D face, the resulting depth ambiguity makes the problem ill-posed. Most existing methods rely on data-driven priors that are built from limited 3D face scans. In contrast, we propose multi-frame video-based self-supervised training of a deep network that (i) learns a face identity model both in shape and appearance while (ii) jointly learning to reconstruct 3D faces. Our face model is learned using only corpora of in-the-wild video clips collected from the Internet. This virtually endless source o

f training data enables learning of a highly general 3D face model. In order to achieve this, we propose a novel multi-frame consistency loss that ensures consistent shape and appearance across multiple frames of a subject's face, thus minimizing depth ambiguity. At test time we can use an arbitrary number of frames, so that we can perform both monocular as well as multi-frame reconstruction.
****

AdaCos: Adaptively Scaling Cosine Logits for Effectively Learning Deep Face Representations
Xiao Zhang, Rui Zhao, Yu Qiao, Xiaogang Wang, Hongsheng Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10823-10832
The cosine-based softmax losses and their variants achieve great success in deep learning based face recognition. However, hyperparameter settings in these losses have significant influences on the optimization path as well as the final recognition performance. Manually tuning those hyperparameters heavily relies on user experience and requires many training tricks. In this paper, we investigate in depth the effects of two important hyperparameters of cosine-based softmax losses, the scale parameter and angular margin parameter, by analyzing how they modulate the predicted classification probability. Based on these analysis, we propose a novel cosine-based softmax loss, AdaCos, which is hyperparameter-free and leverages an adaptive scale parameter to automatically strengthen the training supervisions during the training process. We apply the proposed AdaCos loss to large-scale face verification and identification datasets, including LFW, MegaFace, and IJB-C 1:1 Verification. Our results show that training deep neural networks with the AdaCos loss is stable and able to achieve high face recognition accuracy. Our method outperforms state-of-the-art softmax losses on all the three datasets.
****

3D Hand Shape and Pose Estimation From a Single RGB Image
Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, Junsong Yuan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10833-10842
This work addresses a novel and challenging problem of estimating the full 3D hand shape and pose from a single RGB image. Most current methods in 3D hand analysis from monocular RGB images only focus on estimating the 3D locations of hand keypoints, which cannot fully express the 3D shape of hand. In contrast, we propose a Graph Convolutional Neural Network (Graph CNN) based method to reconstruct a full 3D mesh of hand surface that contains richer information of both 3D hand shape and pose. To train networks with full supervision, we create a large-scale synthetic dataset containing both ground truth 3D meshes and 3D poses. When fine-tuning the networks on real-world datasets without 3D ground truth, we propose a weakly-supervised approach by leveraging the depth map as a weak supervision in training. Through extensive evaluations on our proposed new datasets and two public datasets, we show that our proposed method can produce accurate and reasonable 3D hand mesh, and can achieve superior 3D hand pose estimation accuracy when compared with state-of-the-art methods.
****

3D Hand Shape and Pose From Images in the Wild
Adnane Boukhayma, Rodrigo de Bem, Philip H.S. Torr; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10843-10852
We present in this work the first end-to-end deep learning based method that predicts both 3D hand shape and pose from RGB images in the wild. Our network consists of the concatenation of a deep convolutional encoder, and a fixed model-based decoder. Given an input image, and optionally 2D joint detections obtained from an independent CNN, the encoder predicts a set of hand and view parameters. The decoder has two components: A pre-computed articulated mesh deformation hand model that generates a 3D mesh from the hand parameters, and a re-projection module controlled by the view parameters that projects the generated hand into the image domain. We show that using the shape and pose prior knowledge encoded in th

e hand model within a deep learning framework yields state-of-the-art performanc e in 3D pose prediction from images on standard benchmarks, and produces geometr ically valid and plausible 3D reconstructions. Additionally, we show that traini ng with weak supervision in the form of 2D joint annotations on datasets of imag es in the wild, in conjunction with full supervision in the form of 3D joint ann otations on limited available datasets allows for good generalization to 3D shap e and pose predictions on images in the wild.
****

Self-Supervised 3D Hand Pose Estimation Through Training by Fitting
Chengde Wan, Thomas Probst, Luc Van Gool, Angela Yao; Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 108 53-10862
We present a self-supervision method for 3D hand pose estimation from depth maps . We begin with a neural network initialized with synthesized data and fine-tune it on real but unlabelled depth maps by minimizing a set of data-fitting terms. By approximating the hand surface with a set of spheres, we design a differenti able hand renderer to align estimates by comparing the rendered and input depth maps. In addition, we place a set of priors including a data-driven term to furt her regulate the estimate's kinematic feasibility. Our method makes highly accur ate estimates comparable to current supervised methods which require large amoun ts of labelled training samples, thereby advancing state-of-the-art in unsupervi sed learning for hand pose estimation.
****

CrowdPose: Efficient Crowded Scenes Pose Estimation and a New Benchmark
Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, Cewu Lu; Proceedin gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10863-10872
Multi-person pose estimation is fundamental to many computer vision tasks and ha s made significant progress in recent years. However, few previous methods explo red the problem of pose estimation in crowded scenes while it remains challengin g and inevitable in many scenarios. Moreover, current benchmarks cannot provide an appropriate evaluation for such cases. In this paper, we propose a novel and efficient method to tackle the problem of pose estimation in the crowd and a new dataset to better evaluate algorithms. Our model consists of two key components : joint-candidate single person pose estimation (SPPE) and global maximum joints association. With multi-peak prediction for each joint and global association u sing the graph model, our method is robust to inevitable interference in crowded scenes and very efficient in inference. The proposed method surpasses the stat e-of-the-art methods on CrowdPose dataset by 5.2 mAP and results on MSCOCO datas et demonstrate the generalization ability of our method.
****

Towards Social Artificial Intelligence: Nonverbal Social Signal Prediction in a Triadic Interaction
Hanbyul Joo, Tomas Simon, Mina Cikara, Yaser Sheikh; Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1087 3-10883
We present a new research task and a dataset to understand human social interact ions via computational methods, to ultimately endow machines with the ability to encode and decode a broad channel of social signals humans use. This research d irection is essential to make a machine that genuinely communicates with humans, which we call Social Artificial Intelligence. We first formulate the "social si gnal prediction" problem as a way to model the dynamics of social signals exchan ged among interacting individuals in a data-driven way. We then present a new 3D motion capture dataset to explore this problem, where the broad spectrum of soc ial signals (3D body, face, and hand motions) are captured in a triadic social i nteraction scenario. Baseline approaches to predict speaking status, social form ation, and body gestures of interacting individuals are presented in the defined social prediction framework.
****

HoloPose: Holistic 3D Human Reconstruction In-The-Wild

Riza Alp Guler, Iasonas Kokkinos; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10884-10894

We introduce HoloPose, a method for holistic monocular 3D human body reconstruction. We first introduce a part-based model for 3D model parameter regression that allows our method to operate in-the-wild, gracefully handling severe occlusions and large pose variation. We further train a multi-task network comprising 2D, 3D and Dense Pose estimation to drive the 3D reconstruction task. For this we introduce an iterative refinement method that aligns the model-based 3D estimates of 2D/3D joint positions and DensePose with their image-based counterparts delivered by CNNs, achieving both model-based, global consistency and high spatial accuracy thanks to the bottom-up CNN processing. We validate our contributions on challenging benchmarks, showing that our method allows us to get both accurate joint and 3D surface estimates while operating at more than 10fps in-the-wild. More information about our approach, including videos and demos is available at http://arielai.com/holopose.

****

Weakly-Supervised Discovery of Geometry-Aware Representation for 3D Human Pose Estimation

Xipeng Chen, Kwan-Yee Lin, Wentao Liu, Chen Qian, Liang Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10895-10904

Recent studies have shown remarkable advances in 3D human pose estimation from monocular images, with the help of large-scale in-door 3D datasets and sophisticated network architectures. However, the generalizability to different environments remains an elusive goal. In this work, we propose a geometry-aware 3D representation for the human pose to address this limitation by using multiple views in a simple auto-encoder model at the training stage and only 2D keypoint information as supervision. A view synthesis framework is proposed to learn the shared 3D representation between viewpoints with synthe- sizing the human pose from one viewpoint to the other one. Instead of performing a direct transfer in the raw image- level, we propose a skeleton-based encoder-decoder mechanism to distil only pose-related representation in the latent space. A learning-based representation consistency constraint is further introduced to facilitate the robustness of latent 3D representation. Since the learnt representation encodes 3D geometry information, mapping it to 3D pose will be much easier than conventional frameworks that use an image or 2D coordinates as the input of 3D pose estimator. We demonstrate our approach on the task of 3D human pose estimation. Comprehensive experiments on three popular benchmarks show that our model can significantly improve the performance of state-of-the-art methods with simply injecting the representation as a robust 3D prior.

****

In the Wild Human Pose Estimation Using Explicit 2D Features and Intermediate 3D Representations

Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Gerard Pons-Moll, Christian Theobalt; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10905-10914

Convolutional Neural Network based approaches for monocular 3D human pose estimation usually require a large amount of training images with 3D pose annotations. While it is feasible to provide 2D joint annotations for large corpora of in-the-wild images with humans, providing accurate 3D annotations to such in-the-wild corpora is hardly feasible in practice. Most existing 3D labelled data sets are either synthetically created or feature in-studio images. 3D pose estimation algorithms trained on such data often have limited ability to generalize to real world scene diversity. We therefore propose a new deep learning based method for monocular 3D human pose estimation that shows high accuracy and generalizes better to in-the-wild scenes. It has a network architecture that comprises a new disentangled hidden space encoding of explicit 2D and 3D features, and uses supervision by a new learned projection model from predicted 3D pose. Our algorithm can be jointly trained on image data with 3D labels and image data with only 2D labels. It achieves state-of-the-art accuracy on challenging in-the-wild data.

****
Slim DensePose: Thrifty Learning From Sparse Annotations and Motion Cues
Natalia Neverova, James Thewlis, Riza Alp Guler, Iasonas Kokkinos, Andrea Vedaldi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10915-10923
DensePose supersedes traditional landmark detectors by densely mapping image pixels to body surface coordinates. This power, however, comes at a greatly increased annotation cost, as supervising the model requires to manually label hundreds of points per pose instance. In this work, we thus seek methods to significantly slim down the DensePose annotations, proposing more efficient data collection strategies. In particular, we demonstrate that if annotations are collected in video frames, their efficacy can be multiplied for free by using motion cues. To explore this idea, we introduce DensePose-Track, a dataset of videos where selected frames are annotated in the traditional DensePose manner. Then, building on geometric properties of the DensePose mapping, we use the video dynamic to propagate ground-truth annotations in time as well as to learn from Siamese equivariance constraints. Having performed exhaustive empirical evaluation of various data annotation and learning strategies, we demonstrate that doing so can deliver significantly improved pose estimation results over strong baselines. However, despite what is suggested by some recent works, we show that merely synthesizing motion patterns by applying geometric transformations to isolated frames is significantly less effective, and that motion cues help much more when they are extracted from videos.
****
Self-Supervised Representation Learning From Videos for Facial Action Unit Detection
Yong Li, Jiabei Zeng, Shiguang Shan, Xilin Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10924-10933
In this paper, we aim to learn discriminative representation for facial action unit (AU) detection from large amount of videos without manual annotations. Inspired by the fact that facial actions are the movements of facial muscles, we depict the movements as the transformation between two face images in different frames and use it as the self-supervisory signal to learn the representations. However, under the uncontrolled condition, the transformation is caused by both facial actions and head motions. To remove the influence by head motions, we propose a Twin-Cycle Autoencoder (TCAE) that can disentangle the facial action related movements and the head motion related ones. Specifically, TCAE is trained to respectively change the facial actions and head poses of the source face to those of the target face. Our experiments validate TCAE's capability of decoupling the movements. Experimental results also demonstrate that the learned representation is discriminative for AU detection, where TCAE outperforms or is comparable with the state-of-the-art self-supervised learning methods and supervised AU detection methods.
****
Combining 3D Morphable Models: A Large Scale Face-And-Head Model
Stylianos Ploumpis, Haoyang Wang, Nick Pears, William A. P. Smith, Stefanos Zafeiriou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10934-10943
Three-dimensional Morphable Models (3DMMs) are powerful statistical tools for representing the 3D surfaces of an object class. In this context, we identify an interesting question that has previously not received research attention: is it possible to combine two or more 3DMMs that (a) are built using different templates that perhaps only partly overlap, (b) have different representation capabilities and (c) are built from different datasets that may not be publicly-available? In answering this question, we make two contributions. First, we propose two methods for solving this problem: i. use a regressor to complete missing parts of one model using the other, ii. use the Gaussian Process framework to blend covariance matrices from multiple models. Second, as an example application of our approach, we build a new head and face model that combines the variability and fac

ial detail of the LSFM with the full head modelling of the LYHM. The resulting c
ombined model achieves state-of-the-art performance and outperforms existing hea
d models by a large margin. Finally, as an application experiment, we reconstruc
t full head representations from single, unconstrained images by utilizing our p
roposed large-scale model in conjunction with the Face-Warehouse blendshapes for
 handling expressions.
****

Boosting Local Shape Matching for Dense 3D Face Correspondence
Zhenfeng Fan, Xiyuan Hu, Chen Chen, Silong Peng; Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10944-10
954
Dense 3D face correspondence is a fundamental and challenging issue in the liter
ature of 3D face analysis. Correspondence between two 3D faces can be viewed as
a non-rigid registration problem that one deforms into the other, which is commo
nly guided by a few facial landmarks in many existing works. However, the curren
t works seldom consider the problem of incoherent deformation caused by landmark
s. In this paper, we explicitly formulate the deformation as locally rigid motio
ns guided by some seed points, and the formulated deformation satisfies coherent
 local motions everywhere on a face. The seed points are initialized by a few la
ndmarks, and are then augmented to boost shape matching between the template and
 the target face step by step, to finally achieve dense correspondence. In each
step, we employ a hierarchical scheme for local shape registration, together wit
h a Gaussian reweighting strategy for accurate matching of local features around
 the seed points. In our experiments, we evaluate the proposed method extensivel
y on several datasets, including two publicly available ones: FRGC v2.0 and BU-3
DFE. The experimental results demonstrate that our method can achieve accurate f
eature correspondence, coherent local shape motion, and compact data representat
ion. These merits actually settle some important issues for practical applicatio
ns, such as expressions, noise, and partial data.
****

Unsupervised Part-Based Disentangling of Object Shape and Appearance
Dominik Lorenz, Leonard Bereska, Timo Milbich, Bjorn Ommer; Proceedings of th
e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, p
p. 10955-10964
Large intra-class variation is the result of changes in multiple object characte
ristics. Images, however, only show the superposition of different variable fact
ors such as appearance or shape. Therefore, learning to disentangle and represen
t these different characteristics poses a great challenge, especially in the uns
upervised case. Moreover, large object articulation calls for a flexible part-ba
sed model. We present an unsupervised approach for disentangling appearance and
shape by learning parts consistently over all instances of a category. Our model
 for learning an object representation is trained by simultaneously exploiting i
nvariance and equivariance constraints between synthetically transformed images.
 Since no part annotation or prior information on an object class is required, t
he approach is applicable to arbitrary classes. We evaluate our approach on a wi
de range of object categories and diverse tasks including pose prediction, disen
tangled image synthesis, and video-to-video translation. The approach outperform
s the state-of-the-art on unsupervised keypoint prediction and compares favorabl
y even against supervised approaches on the task of shape and appearance transfe
r.
****

Monocular Total Capture: Posing Face, Body, and Hands in the Wild
Donglai Xiang, Hanbyul Joo, Yaser Sheikh; Proceedings of the IEEE/CVF Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10965-10974
We present the first method to capture the 3D total motion of a target person fr
om a monocular view input. Given an image or a monocular video, our method recon
structs the motion from body, face, and fingers represented by a 3D deformable m
esh model. We use an efficient representation called 3D Part Orientation Fields
(POFs), to encode the 3D orientations of all body parts in the common 2D image s
pace. POFs are predicted by a Fully Convolutional Network, along with the joint

confidence maps. To train our network, we collect a new 3D human motion dataset capturing diverse total body motion of 40 subjects in a multiview system. We leverage a 3D deformable human model to reconstruct total body pose from the CNN outputs with the aid of the pose and shape prior in the model. We also present a texture-based tracking method to obtain temporally coherent motion capture output. We perform thorough quantitative evaluations including comparison with the existing body-specific and hand-specific methods, and performance analysis on camera viewpoint and human pose changes. Finally, we demonstrate the results of our total body motion capture on various challenging in-the-wild videos.
****

Expressive Body Capture: 3D Hands, Face, and Body From a Single Image
Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, Michael J. Black; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10975-10985
To facilitate the analysis of human actions, interactions and emotions, we compute a 3D model of human body pose, hand pose, and facial expression from a single monocular image. To achieve this, we use thousands of 3D scans to train a new, unified, 3D model of the human body, SMPL-X, that extends SMPL with fully articulated hands and an expressive face. Learning to regress the parameters of SMPL-X directly from images is challenging without paired images and 3D ground truth. Consequently, we follow the approach of SMPLify, which estimates 2D features and then optimizes model parameters to fit the features. We improve on SMPLify in several significant ways: (1) we detect 2D features corresponding to the face, hands, and feet and fit the full SMPL-X model to these; (2) we train a new neural network pose prior using a large MoCap dataset; (3) we define a new interpenetration penalty that is both fast and accurate; (4) we automatically detect gender and the appropriate body models (male, female, or neutral); (5) our PyTorch implementation achieves a speedup of more than 8x over Chumpy. We use the new method, SMPLify-X, to fit SMPL-X to both controlled images and images in the wild. We evaluate 3D accuracy on a new curated dataset comprising 100 images with pseudo ground-truth. This is a step towards automatic expressive human capture from monocular RGB data. The models, code, and data are available for research purposes at https://smpl-x.is.tue.mpg.de.
****

Neural RGB(r)D Sensing: Depth and Uncertainty From a Video Camera
Chao Liu, Jinwei Gu, Kihwan Kim, Srinivasa G. Narasimhan, Jan Kautz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10986-10995
Depth sensing is crucial for 3D reconstruction and scene understanding. Active depth sensors provide dense metric measurements, but often suffer from limitations such as restricted operating ranges, low spatial resolution, sensor interference, and high power consumption. In this paper, we propose a deep learning (DL) method to estimate per-pixel depth and its uncertainty continuously from a monocular video stream, with the goal of effectively turning an RGB camera into an RGB-D camera. Unlike prior DL-based methods, we estimate a depth probability distribution for each pixel rather than a single depth value, leading to an estimate of a 3D depth probability volume for each input frame. These depth probability volumes are accumulated over time under a Bayesian filtering framework as more incoming frames are processed sequentially, which effectively reduces depth uncertainty and improves accuracy, robustness, and temporal stability. Compared to prior work, the proposed approach achieves more accurate and stable results, and generalizes better to new datasets. Experimental results also show the output of our approach can be directly fed into classical RGB-D based 3D scanning methods for 3D scene reconstruction.
****

DAVANet: Stereo Deblurring With View Aggregation
Shangchen Zhou, Jiawei Zhang, Wangmeng Zuo, Haozhe Xie, Jinshan Pan, Jimmy S. Ren; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10996-11005
Nowadays stereo cameras are more commonly adopted in emerging devices such as du

al-lens smartphones and unmanned aerial vehicles. However, they also suffer from blurry images in dynamic scenes which leads to visual discomfort and hampers further image processing. Previous works have succeeded in monocular deblurring, yet there are few studies on deblurring for stereoscopic images. By exploiting the two-view nature of stereo images, we propose a novel stereo image deblurring network with Depth Awareness and View Aggregation, named DAVANet. In our proposed network, 3D scene cues from the depth and varying information from two views are incorporated, which help to remove complex spatially-varying blur in dynamic scenes. Specifically, with our proposed fusion network, we integrate the bidirectional disparities estimation and deblurring into a unified framework. Moreover, we present a large-scale multi-scene dataset for stereo deblurring, containing 20,637 blurry-sharp stereo image pairs from 135 diverse sequences and their corresponding bidirectional disparities. The experimental results on our dataset demonstrate that DAVANet outperforms state-of-the-art methods in terms of accuracy, speed, and model size.

****

DVC: An End-To-End Deep Video Compression Framework

Guo Lu, Wanli Ouyang, Dong Xu, Xiaoyun Zhang, Chunlei Cai, Zhiyong Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11006-11015

Conventional video compression approaches use the predictive coding architecture and encode the corresponding motion information and residual information. In this paper, taking advantage of both classical architecture in the conventional video compression method and the powerful non-linear representation ability of neural networks, we propose the first end-to-end video compression deep model that jointly optimizes all the components for video compression. Specifically, learning based optical flow estimation is utilized to obtain the motion information and reconstruct the current frames. Then we employ two auto-encoder style neural networks to compress the corresponding motion and residual information. All the modules are jointly learned through a single loss function, in which they collaborate with each other by considering the trade-off between reducing the number of compression bits and improving quality of the decoded video. Experimental results show that the proposed approach can outperform the widely used video coding standard H.264 in terms of PSNR and be even on par with the latest standard H.265 in terms of MS-SSIM. Code is released at https://github.com/GuoLusjtu/DVC.

****

SOSNet: Second Order Similarity Regularization for Local Descriptor Learning

Yurun Tian, Xin Yu, Bin Fan, Fuchao Wu, Huub Heijnen, Vassileios Balntas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11016-11025

Despite the fact that Second Order Similarity (SOS) has been used with significant success in tasks such as graph matching and clustering, it has not been exploited for learning local descriptors. In this work, we explore the potential of \sos in the field of descriptor learning by building upon the intuition that a positive pair of matching points should exhibit similar distances with respect to other points in the embedding space. Thus, we propose a novel regularization term, named Second Order Similarity Regularization (SOSR), that follows this principle. By incorporating SOSR into training, our learned descriptor achieves state-of-the-art performance on several challenging benchmarks containing distinct tasks ranging from local patch retrieval to structure from motion. Furthermore, by designing a von Mises-Fischer distribution based evaluation method, we link the utilization of the descriptor space to the matching performance, thus demonstrating the effectiveness of our proposed SOSR. Extensive experimental results, empirical evidence, and in-depth analysis are provided, indicating that SOSR can significantly boost the matching performance of the learned descriptor.

****

"Double-DIP": Unsupervised Image Decomposition via Coupled Deep-Image-Priors

Yosef Gandelsman, Assaf Shocher, Michal Irani; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11026-11035

Many seemingly unrelated computer vision tasks can be viewed as a special case o

f image decomposition into separate layers. For example, image segmentation (se
paration into foreground and background layers); transparent layer separation (i
nto reflection and transmission layers); Image dehazing (separation into a clear
image and a haze map), and more. In this paper we propose a unified framework f
or unsupervised layer decomposition of a single image, based on coupled "Deep-im
age-Prior" (DIP) networks. It was shown [Ulyanov et al] that the structure of a
single DIP generator network is sufficient to capture the low-level statistics
of a single image. We show that coupling multiple such DIPs provides a powerful
tool for decomposing images into their basic components, for a wide variety of a
pplications. This capability stems from the fact that the internal statistics of
a mixture of layers is more complex than the statistics of each of its individu
al components. We show the power of this approach for Image-Dehazing, Fg/Bg Segm
entation, Watermark-Removal, Transparency Separation in images and video, and mo
re. These capabilities are achieved in a totally unsupervised way, with no train
ing examples other than the input image/video itself.
****

Unprocessing Images for Learned Raw Denoising
Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, Jonat
han T. Barron; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat
tern Recognition (CVPR), 2019, pp. 11036-11045
Machine learning techniques work best when the data used for training resembles
the data used for evaluation. This holds true for learned single-image denoising
algorithms, which are applied to real raw camera sensor readings but, due to pr
actical constraints, are often trained on synthetic image data. Though it is und
erstood that generalizing from synthetic to real images requires careful conside
ration of the noise properties of camera sensors, the other aspects of an image
processing pipeline (such as gain, color correction, and tone mapping) are often
overlooked, despite their significant effect on how raw measurements are transf
ormed into finished images. To address this, we present a technique to "unproces
s" images by inverting each step of an image processing pipeline, thereby allowi
ng us to synthesize realistic raw sensor measurements from commonly available In
ternet photos. We additionally model the relevant components of an image process
ing pipeline when evaluating our loss function, which allows training to be awar
e of all relevant photometric processing that will occur after denoising. By unp
rocessing and processing training data and model outputs in this way, we are abl
e to train a simple convolutional neural network that has 14%-38% lower error ra
tes and is 9x-18x faster than the previous state of the art on the Darmstadt Noi
se Dataset, and generalizes to sensors outside of that dataset as well.
****

Residual Networks for Light Field Image Super-Resolution
Shuo Zhang, Youfang Lin, Hao Sheng; Proceedings of the IEEE/CVF Conference on
Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11046-11055
Light field cameras are considered to have many potential applications since ang
ular and spatial information is captured simultaneously. However, the limited sp
atial resolution has brought lots of difficulties in developing related applicat
ions and becomes the main bottleneck of light field cameras. In this paper, a le
arning-based method using residual convolutional networks is proposed to reconst
ruct light fields with higher spatial resolution. The view images in one light f
ield are first grouped into different image stacks with consistent sub-pixel off
sets and fed into different network branches to implicitly learn inherent corres
ponding relations. The residual information in different spatial directions is t
hen calculated from each branch and further integrated to supplement high-freque
ncy details for the view image. Finally, a flexible solution is proposed to sup
er-resolve entire light field images with various angular resolutions. Experimen
tal results on synthetic and real-world datasets demonstrate that the proposed m
ethod outperforms other state-of-the-art methods by a large margin in both visua
l and numerical evaluations. Furthermore, the proposed method shows good perform
ances in preserving the inherent epipolar property in light field images.
****

Modulating Image Restoration With Continual Levels via Adaptive Feature Modifica

tion Layers
Jingwen He, Chao Dong, Yu Qiao; Proceedings of the IEEE/CVF Conference on Comp uter Vision and Pattern Recognition (CVPR), 2019, pp. 11056-11064
In image restoration tasks, like denoising and superresolution, continual modula tion of restoration levels is of great importance for real-world applications, b ut has failed most of existing deep learning based image restoration methods. Le arning from discrete and fixed restoration levels, deep models cannot be easily generalized to data of continuous and unseen levels. This topic is rarely touche d in literature, due to the difficulty of modulating well-trained models with ce rtain hyper-parameters. We make a step forward by proposing a unified CNN framew ork that consists of little additional parameters than a single-level model yet could handle arbitrary restoration levels between a start and an end level. The additional module, namely AdaFM layer, performs channel-wise feature modificatio n, and can adapt a model to another restoration level with high accuracy. By sim ply tweaking an interpolation coefficient, the intermediate model - AdaFM-Net co uld generate smooth and continuous restoration effects without artifacts. Extens ive experiments on three image restoration tasks demonstrate the effectiveness o f both model training and modulation testing. Besides, we carefully investigate the properties of AdaFM layers, providing a detailed guidance on the usage of th e proposed method.
****

Second-Order Attention Network for Single Image Super-Resolution
Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, Lei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019 , pp. 11065-11074
Recently, deep convolutional neural networks (CNNs) have been widely explored in single image super-resolution (SISR) and obtained remarkable performance. Howev er, most of the existing CNN-based SISR methods mainly focus on wider or deeper architecture design, neglecting to explore the feature correlations of intermedi ate layers, hence hindering the representational power of CNNs. To address this issue, in this paper, we propose a second-order attention network (SAN) for more powerful feature expression and feature correlation learning. Specifically, a n ovel train- able second-order channel attention (SOCA) module is developed to ad aptively rescale the channel-wise features by using second-order feature statist ics for more discriminative representations. Furthermore, we present a non-local ly enhanced residual group (NLRG) structure, which not only incorporates non-loc al operations to capture long-distance spatial contextual information, but also contains repeated local-source residual attention groups (LSRAG) to learn increa singly abstract feature representations. Experimental results demonstrate the su periority of our SAN network over state-of-the-art SISR methods in terms of both quantitative metrics and visual quality.
****

Devil Is in the Edges: Learning Semantic Boundaries From Noisy Annotations
David Acuna, Amlan Kar, Sanja Fidler; Proceedings of the IEEE/CVF Conference o n Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11075-11083
We tackle the problem of semantic boundary prediction, which aims to identify pi xels that belong to object(class) boundaries. We notice that relevant datasets c onsist of a significant level of label noise, reflecting the fact that precise a nnotations are laborious to get and thus annotators trade-off quality with effic iency. We aim to learn sharp and precise semantic boundaries by explicitly reaso ning about annotation noise during training. We propose a simple new layer and l oss that can be used with existing learning-based boundary detectors. Our layer/ loss enforces the detector to predict a maximum response along the normal direct ion at an edge, while also regularizing its direction. We further reason about t rue object boundaries during training using a level set formulation, which allow s the network to learn from misaligned labels in an end-to-end fashion. Experime nts show that we improve over the CASENet backbone network by more than 4% in te rms of MF(ODS) and 18.61% in terms of AP, outperforming all current state-of-the -art methods including those that deal with alignment. Furthermore, we show that our learned network can be used to significantly improve coarse segmentation la

bels, lending itself as an efficient way to label new data.
****

Path-Invariant Map Networks
Zaiwei Zhang, Zhenxiao Liang, Lemeng Wu, Xiaowei Zhou, Qixing Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11084-11094
Optimizing a network of maps among a collection of objects/domains (or map synchronization) is a central problem across computer vision and many other relevant fields. Compared to optimizing pairwise maps in isolation, the benefit of map synchronization is that there are natural constraints among a map network that can improve the quality of individual maps. While such self-supervision constraints are well-understood for undirected map networks (e.g., the cycle-consistency constraint), they are under-explored for directed map networks, which naturally arise when maps are given by parametric maps (e.g., a feed-forward neural network). In this paper, we study a natural self-supervision constraint for directed map networks called path-invariance, which enforces that composite maps along different paths between a fixed pair of source and target domains are identical. We introduce path-invariance bases for efficient encoding of the path-invariance constraint and present an algorithm that outputs a path-variance basis with polynomial time and space complexities. We demonstrate the effectiveness of our formulation on optimizing object correspondences, estimating dense image maps via neural networks, and 3D scene segmentation via map networks of diverse 3D representations. In particular, our approach only requires 8% labeled data from ScanNet to achieve the same performance as training a single 3D semantic segmentation network with 30% to 100% labeled data.
****

FilterReg: Robust and Efficient Probabilistic Point-Set Registration Using Gaussian Filter and Twist Parameterization
Wei Gao, Russ Tedrake; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11095-11104
Probabilistic point-set registration methods have been gaining more attention for their robustness to noise, outliers and occlusions. However, these methods tend to be much slower than the popular iterative closest point (ICP) algorithms, which severely limits their usability. In this paper, we contribute a novel probabilistic registration method that achieves state-of-the-art robustness as well as substantially faster computational performance than modern ICP implementations. This is achieved using a rigorous yet computationally-efficient probabilistic formulation. Point-set registration is cast as a maximum likelihood estimation and solved using the EM algorithm. We show that with a simple augmentation, the E step can be formulated as a filtering problem, allowing us to leverage advances in efficient Gaussian filtering methods. We also propose a customized permutohedral filter to improve its performance while retaining sufficient accuracy for our task. Additionally, we present a simple and efficient twist parameterization that generalizes our method to the registration of articulated and deformable objects. For articulated objects, the complexity of our method is almost independent of the Degrees Of Freedom (DOFs), which makes it highly efficient even for high DOF systems. The results demonstrate the proposed method consistently outperforms many competitive baselines on a variety of registration tasks.
****

Probabilistic Permutation Synchronization Using the Riemannian Structure of the Birkhoff Polytope
Tolga Birdal, Umut Simsekli; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11105-11116
We present an entirely new geometric and probabilistic approach to synchronization of correspondences across multiple sets of objects or images. In particular, we present two algorithms: (1) Birkhoff-Riemannian L-BFGS for optimizing the relaxed version of the combinatorially intractable cycle consistency loss in a principled manner, (2) Birkhoff-Riemannian Langevin Monte Carlo for generating samples on the Birkhoff Polytope and estimating the confidence of the found solutions. To this end, we first introduce the very recently developed Riemannian geometr

y of the Birkhoff Polytope. Next, we introduce a new probabilistic synchronizati on model in the form of a Markov Random Field (MRF). Finally, based on the first order retraction operators, we formulate our problem as simulating a stochastic differential equation and devise new integrators. We show on both synthetic and real datasets that we achieve high quality multi-graph matching results with fa ster convergence and reliable confidence/uncertainty estimates.
****

Lifting Vectorial Variational Problems: A Natural Formulation Based on Geometric Measure Theory and Discrete Exterior Calculus
Thomas Mollenhoff, Daniel Cremers; Proceedings of the IEEE/CVF Conference on Co mputer Vision and Pattern Recognition (CVPR), 2019, pp. 11117-11126
Numerous tasks in imaging and vision can be formulated as variational problems o ver vector-valued maps. We approach the relaxation and convexification of such vectorial variational problems via a lifting to the space of currents. To that e nd, we recall that functionals with polyconvex Lagrangians can be reparametrized as convex one-homogeneous functionals on the graph of the function. This leads to an equivalent shape optimization problem over oriented surfaces in the produc t space of domain and codomain. A convex formulation is then obtained by relaxin g the search space from oriented surfaces to more general currents. We propose a discretization of the resulting infinite-dimensional optimization problem using Whitney forms, which also generalizes recent "sublabel-accurate" multilabeling approaches.
****

A Sufficient Condition for Convergences of Adam and RMSProp
Fangyu Zou, Li Shen, Zequn Jie, Weizhong Zhang, Wei Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11127-11135
Adam and RMSProp are two of the most influential adaptive stochastic algorithms for training deep neural networks, which have been pointed out to be divergent e ven in the convex setting via a few simple counterexamples. Many attempts, such as decreasing an adaptive learning rate, adopting a big batch size, incorporatin g a temporal decorrelation technique, seeking an analogous surrogate, etc., have been tried to promote Adam/RMSProp-type algorithms to converge. In contrast wit h existing approaches, we introduce an alternative easy-to-check sufficient cond ition, which merely depends on the parameters of the base learning rate and comb inations of historical second-order moments, to guarantee the global convergence of generic Adam/RMSProp for solving large-scale non-convex stochastic optimizat ion. Moreover, we show that the convergences of several variants of Adam, such a s AdamNC, AdaEMA, etc., can be directly implied via the proposed sufficient cond ition in the non-convex setting. In addition, we illustrate that Adam is essenti ally a specifically weighted AdaGrad with exponential moving average momentum, w hich provides a novel perspective for understanding Adam and RMSProp. This obser vation coupled with this sufficient condition gives much deeper interpretations on their divergences. At last, we validate the sufficient condition by applying Adam and RMSProp to tackle a certain counterexample and train deep neural networ ks. Numerical results are exactly in accord with our theoretical analysis.
****

Guaranteed Matrix Completion Under Multiple Linear Transformations
Chao Li, Wei He, Longhao Yuan, Zhun Sun, Qibin Zhao; Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 111 36-11145
Low-rank matrix completion (LRMC) is a classical model in both computer vision (CV) and machine learning, and has been successfully applied to various real appl ications. In the recent CV tasks, the completion is usually employed on the vari ants of data, such as "non-local" or filtered, rather than their original forms. This fact makes that the theoretical analysis of the conventional LRMC is no lo nger suitable in these applications. To tackle this problem, we propose a more g eneral framework for LRMC, in which the linear transformations of the data are t aken into account. We rigorously prove the identifiability of the proposed model and show an upper bound of the reconstruction error. Furthermore, we derive an

efficient completion algorithm by using augmented Lagrangian multipliers and the sketching trick. In the experiments, we apply the proposed method to the classical image inpainting problem and achieve the state-of-the-art results.
****

MAP Inference via Block-Coordinate Frank-Wolfe Algorithm
Paul Swoboda, Vladimir Kolmogorov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11146-11155
We present a new proximal bundle method for Maximum-A-Posteriori (MAP) inference in structured energy minimization problems. The method optimizes a Lagrangean relaxation of the original energy minimization problem using a multi plane block-coordinate Frank-Wolfe method that takes advantage of the specific structure of the Lagrangean decomposition. We show empirically that our method outperforms state-of-the-art Lagrangean decomposition based algorithms on some challenging Markov Random Field, multi-label discrete tomography and graph matching problems.
****

A Convex Relaxation for Multi-Graph Matching
Paul Swoboda, Dagmar Kainm"uller, Ashkan Mokarian, Christian Theobalt, Florian Bernard; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11156-11165
We present a convex relaxation for the multi-graph matching problem. Our formulation allows for partial pairwise matchings, guarantees cycle consistency, and our objective incorporates both linear and quadratic costs. Moreover, we also present an extension to higher-order costs. In order to solve the convex relaxation we employ a message passing algorithm that optimizes the dual problem. We experimentally compare our algorithm on established benchmark problems from computer vision, as well as on large problems from biological image analysis, the size of which exceed previously investigated multi-graph matching instances.
****

Pixel-Adaptive Convolutional Neural Networks
Hang Su, Varun Jampani, Deqing Sun, Orazio Gallo, Erik Learned-Miller, Jan Kautz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11166-11175
Convolutions are the fundamental building blocks of CNNs. The fact that their weights are spatially shared is one of the main reasons for their widespread use, but it is also a major limitation, as it makes convolutions content-agnostic. We propose a pixel-adaptive convolution (PAC) operation, a simple yet effective modification of standard convolutions, in which the filter weights are multiplied with a spatially varying kernel that depends on learnable, local pixel features. PAC is a generalization of several popular filtering techniques and thus can be used for a wide range of use cases. Specifically, we demonstrate state-of-the-art performance when PAC is used for deep joint image upsampling. PAC also offers an effective alternative to fully-connected CRF (Full-CRF), called PAC-CRF, which performs competitively compared to Full-CRF, while being considerably faster. In addition, we also demonstrate that PAC can be used as a drop-in replacement for convolution layers in pre-trained networks, resulting in consistent performance improvements.
****

Single-Frame Regularization for Temporally Stable CNNs
Gabriel Eilertsen, Rafal K. Mantiuk, Jonas Unger; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11176-11185
Convolutional neural networks (CNNs) can model complicated non-linear relations between images. However, they are notoriously sensitive to small changes in the input. Most CNNs trained to describe image-to-image mappings generate temporally unstable results when applied to video sequences, leading to flickering artifacts and other inconsistencies over time. In order to use CNNs for video material, previous methods have relied on estimating dense frame-to-frame motion information (optical flow) in the training and/or the inference phase, or by exploring recurrent learning structures. We take a different approach to the problem, posing temporal stability as a regularization of the cost function. The regularizatio

n is formulated to account for different types of motion that can occur between frames, so that temporally stable CNNs can be trained without the need for video material or expensive motion estimation. The training can be performed as a fine-tuning operation, without architectural modifications of the CNN. Our evaluation shows that the training strategy leads to large improvements in temporal smoothness. Moreover, for small datasets the regularization can help in boosting the generalization performance to a much larger extent than what is possible with naive augmentation strategies.
****

An End-To-End Network for Generating Social Relationship Graphs
Arushi Goel, Keng Teck Ma, Cheston Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11186-11195
Socially-intelligent agents are of growing interest in artificial intelligence. To this end, we need systems that can understand social relationships in diverse social contexts. Inferring the social context in a given visual scene not only involves recognizing objects, but also demands a more in-depth understanding of the relationships and attributes of the people involved. To achieve this, one computational approach for representing human relationships and attributes is to use an explicit knowledge graph, which allows for high-level reasoning. We introduce a novel end-to-end-trainable neural network that is capable of generating a Social Relationship Graph - a structured, unified representation of social relationships and attributes - from a given input image. Our Social Relationship Graph Generation Network (SRG-GN) is the first to use memory cells like Gated Recurrent Units (GRUs) to iteratively update the social relationship states in a graph using scene and attribute context. The neural network exploits the recurrent connections among the GRUs to implement message passing between nodes and edges in the graph, and results in significant improvement over previous methods for social relationship recognition.
****

Meta-Learning Convolutional Neural Architectures for Multi-Target Concrete Defect Classification With the COncrete DEfect BRidge IMage Dataset
Martin Mundt, Sagnik Majumder, Sreenivas Murali, Panagiotis Panetsos, Visvanathan Ramesh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11196-11205
Recognition of defects in concrete infrastructure, especially in bridges, is a costly and time consuming crucial first step in the assessment of the structural integrity. Large variation in appearance of the concrete material, changing illumination and weather conditions, a variety of possible surface markings as well as the possibility for different types of defects to overlap, make it a challenging real-world task. In this work we introduce the novel COncrete DEfect BRidge IMage dataset (CODEBRIM) for multi-target classification of five commonly appearing concrete defects. We investigate and compare two reinforcement learning based meta-learning approaches, MetaQNN and efficient neural architecture search, to find suitable convolutional neural network architectures for this challenging multi-class multi-target task. We show that learned architectures have fewer overall parameters in addition to yielding better multi-target accuracy in comparison to popular neural architectures from the literature evaluated in the context of our application.
****

ECC: Platform-Independent Energy-Constrained Deep Neural Network Compression via a Bilinear Regression Model
Haichuan Yang, Yuhao Zhu, Ji Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11206-11215
Many DNN-enabled vision applications constantly operate under severe energy constraints such as unmanned aerial vehicles, Augmented Reality headsets, and smartphones. Designing DNNs that can meet a stringent energy budget is becoming increasingly important. This paper proposes ECC, a framework that compresses DNNs to meet a given energy constraint while minimizing accuracy loss. The key idea of ECC is to model the DNN energy consumption via a novel bilinear regression function. The energy estimate model allows us to formulate DNN compression as a constra

ined optimization that minimizes the DNN loss function over the energy constrain t. The optimization problem, however, has nontrivial constraints. Therefore, exi sting deep learning solvers do not apply directly. We propose an optimization al gorithm that combines the essence of the Alternating Direction Method of Multipl iers (ADMM) framework with gradient-based learning algorithms. The algorithm dec omposes the original constrained optimization into several subproblems that are solved iteratively and efficiently. ECC is also portable across different hardwa re platforms without requiring hardware knowledge. Experiments show that ECC ach ieves higher accuracy under the same or lower energy budget compared to state-of -the-art resource-constrained DNN compression techniques.
****

SeerNet: Predicting Convolutional Neural Network Feature-Map Sparsity Through Lo w-Bit Quantization
Shijie Cao, Lingxiao Ma, Wencong Xiao, Chen Zhang, Yunxin Liu, Lintao Zhang , Lanshun Nie, Zhi Yang; Proceedings of the IEEE/CVF Conference on Computer Vi sion and Pattern Recognition (CVPR), 2019, pp. 11216-11225
In this paper we present a novel and general method to accelerate convolutional neural network (CNN) inference by taking advantage of feature map sparsity. We e xperimentally demonstrate that a highly quantized version of the original networ k is sufficient in predicting the output sparsity accurately, and verify that le veraging such sparsity in inference incurs negligible accuracy drop compared wit h the original network. To accelerate inference, for each convolution layer our approach first obtains a binary sparsity mask of the output feature maps by runn ing inference on a quantized version of the original network layer, and then con ducts a full-precision sparse convolution to find out the precise values of the non-zero outputs. Compared with existing work, our approach avoids the overhead of training additional auxiliary networks, while is still applicable to general CNN networks without being limited to certain application domains.
****

Defending Against Adversarial Attacks by Randomized Diversification
Olga Taran, Shideh Rezaeifar, Taras Holotyak, Slava Voloshynovskiy; Proceedin gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11226-11233
The vulnerability of machine learning systems to adversarial attacks questions t heir usage in many applications. In this paper, we propose a randomized diversif ication as a defense strategy. We introduce a multi-channel architecture in a gr ay-box scenario, which assumes that the architecture of the classifier and the t raining data set are known to the attacker. The attacker does not only have acce ss to a secret key and to the internal states of the system at the test time. Th e defender processes an input in multiple channels. Each channel introduces its own randomization in a special transform domain based on a secret key shared bet ween the training and testing stages. Such a transform based randomization with a shared key preserves the gradients in key-defined sub-spaces for the defender but it prevents gradient back propagation and the creation of various bypass sys tems for the attacker. An additional benefit of multi-channel randomization is t he aggregation that fuses soft-outputs from all channels, thus increasing the re liability of the final score. The sharing of a secret key creates an information advantage to the defender. Experimental evaluation demonstrates an increased ro bustness of the proposed method to a number of known state-of-the-art attacks.
****

Rob-GAN: Generator, Discriminator, and Adversarial Attacker
Xuanqing Liu, Cho-Jui Hsieh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11234-11243
We study two important concepts in adversarial deep learning---adversarial train ing and generative adversarial network (GAN). Adversarial training is the techn ique used to improve the robustness of discriminator by combining adversarial at tacker and discriminator in the training phase. GAN is commonly used for image generation by jointly optimizing discriminator and generator. We show these two concepts are indeed closely related and can be used to strengthen each other---a dding a generator to the adversarial training procedure can improve the robustne

ss of discriminators, and adding an adversarial attack to GAN training can impro ve the convergence speed and lead to better generators. Combining these two insi ghts, we develop a framework called Rob-GAN to jointly optimize generator and di scriminator in the presence of adversarial attacks---the generator generates fak e images to fool discriminator; the adversarial attacker perturbs real images to fool discriminator, and the discriminator wants to minimize loss under fake and adversarial images. Through this end-to-end training procedure, we are able to simultaneously improve the convergence speed of GAN training, the quality of syn thetic images, and the robustness of discriminator under strong adversarial atta cks. Experimental results demonstrate that the obtained classifier is more robus t than the state-of-the-art adversarial training approach (Madry 2017), and the generator outperforms SN-GAN on ImageNet-143.

****

Learning From Noisy Labels by Regularized Estimation of Annotator Confusion

Ryutaro Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Daniel C. Alexander, Nathan Silberman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11244-11253

The predictive performance of supervised learning algorithms depends on the qual ity of labels. In a typical label collection process, multiple annotators provid e subjective noisy estimates of the "truth" under the influence of their varying skill-levels and biases. Blindly treating these noisy labels as the ground trut h limits the accuracy of learning algorithms in the presence of strong disagreem ent. This problem is critical for applications in domains such as medical imagin g where both the annotation cost and inter-observer variability are high. In thi s work, we present a method for simultaneously learning the individual annotator model and the underlying true label distribution, using only noisy observations . Each annotator is modeled by a confusion matrix that is jointly estimated alon g with the classifier predictions. We propose to add a regularization term to th e loss function that encourages convergence to the true annotator confusion matr ix. We provide a theoretical argument as to how the regularization is essential to our approach both for the case of single annotator and multiple annotators. D espite the simplicity of the idea, experiments on image classification tasks wit h both simulated and real labels show that our method either outperforms or perf orms on par with the state-of-the-art methods and is capable of estimating the s kills of annotators even with a single label available per image.

****

Task-Free Continual Learning

Rahaf Aljundi, Klaas Kelchtermans, Tinne Tuytelaars; Proceedings of the IEEE/C VF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11254 -11263

Methods proposed in the literature towards continual deep learning typically op erate in a task-based sequential learning setup. A sequence of tasks is learned, one at a time, with all data of current task available but not of previous or f uture tasks. Task boundaries and identities are known at all times. This setup, however, is rarely encountered in practical applications. Therefore we investig ate how to transform continual learning to an online setup. We develop a system that keeps on learning over time in a streaming fashion, with data distribution s gradually changing and without the notion of separate tasks. To this end, we build on the work on Memory Aware Synapses, and show how this method can be made online by providing a protocol to decide i) when to update the importance weigh ts, ii) which data to use to update them, and iii) how to accumulate the importa nce weights at each update step. Experimental results show the validity of the approach in the context of two applications: (self-)supervised learning of a fac e recognition model by watching soap series and learning a robot to avoid collis ions.

****

Importance Estimation for Neural Network Pruning

Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, Jan Kautz; Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11264-11272

Structural pruning of neural network parameters reduces computational, energy, a
nd memory transfer costs during inference. We propose a novel method that estima
tes the contribution of a neuron (filter) to the final loss and iteratively remo
ves those with smaller scores. We describe two variations of our method using th
e first and second-order Taylor expansions to approximate a filter's contributio
n. Both methods scale consistently across any network layer without requiring pe
r-layer sensitivity analysis and can be applied to any kind of layer, including
skip connections. For modern networks trained on ImageNet, we measured experimen
tally a high (>93%) correlation between the contribution computed by our methods
and a reliable estimate of the true importance. Pruning with the proposed meth
ods led to an improvement over state-of-the-art in terms of accuracy, FLOPs, and
parameter reduction. On ResNet-101, we achieve a 40% FLOPS reduction by removin
g 30% of the parameters, with a loss of 0.02% in the top-1 accuracy on ImageNet.
****

Detecting Overfitting of Deep Generative Networks via Latent Recovery
Ryan Webster, Julien Rabin, Loic Simon, Frederic Jurie; Proceedings of the IE
EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1
1273-11282
State of the art deep generative networks have achieved such realism that they c
an be suspected of memorizing training images. It is why it is not uncommon to i
nclude visualizations of training set nearest neighbors, to suggest generated im
ages are not simply memorized. We argue this is not sufficient and motivates stu
dying overfitting of deep generators with more scrutiny. We address this questio
n by i) showing how simple losses are highly effective at reconstructing images
for deep generators ii) analyzing the statistics of reconstruction errors for tr
aining versus validation images. Using this methodology, we show that pure GAN m
odels appear to generalize well, in contrast with those using hybrid adversarial
losses, which are amongst the most widely applied generative methods. We also s
how that standard GAN evaluation metrics fail to capture memorization for some d
eep generators. Finally, we note the ramifications of memorization on data priva
cy. Considering the already widespread application of generative networks, we pr
ovide a step in the right direction towards the important yet incomplete picture
of generative overfitting.
****

Coloring With Limited Data: Few-Shot Colorization via Memory Augmented Networks
Seungjoo Yoo, Hyojin Bahng, Sunghyo Chung, Junsoo Lee, Jaehyuk Chang, Jaegu
l Choo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Re
cognition (CVPR), 2019, pp. 11283-11292
Despite recent advancements in deep learning-based automatic colorization, they
are still limited when it comes to few-shot learning. Existing models require a
significant amount of training data. To tackle this issue, we present a novel me
mory-augmented colorization model MemoPainter that can produce high-quality colo
rization with limited data. In particular, our model is able to capture rare ins
tances and successfully colorize them. Also, we propose a novel threshold triple
t loss that enables unsupervised training of memory networks without the need fo
r class labels. Experiments show that our model has superior quality in both few
-shot and one-shot colorization tasks.
****

Characterizing and Avoiding Negative Transfer
Zirui Wang, Zihang Dai, Barnabas Poczos, Jaime Carbonell; Proceedings of the
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp.
11293-11302
When labeled data is scarce for a specific target task, transfer learning often
offers an effective solution by utilizing data from a related source task. Howev
er, when transferring knowledge from a less related source, it may inversely hur
t the target performance, a phenomenon known as negative transfer. Despite its p
ervasiveness, negative transfer is usually described in an informal manner, lack
ing rigorous definition, careful analysis, or systematic treatment. This paper p
roposes a formal definition of negative transfer and analyzes three important as
pects thereof. Stemming from this analysis, a novel technique is proposed to cir

cumvent negative transfer by filtering out unrelated source data. Based on adver sarial networks, the technique is highly generic and can be applied to a wide ra nge of transfer learning algorithms. The proposed approach is evaluated on six s tate-of-the-art deep transfer methods via experiments on four benchmark datasets with varying levels of difficulty. Empirically, the proposed method consistentl y improves the performance of all baseline methods and largely avoids negative t ransfer, even when the source data is degenerate.
****

Building Efficient Deep Neural Networks With Unitary Group Convolutions
Ritchie Zhao, Yuwei Hu, Jordan Dotzel, Christopher De Sa, Zhiru Zhang; Proce edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV PR), 2019, pp. 11303-11312
We propose unitary group convolutions (UGConvs), a building block for CNNs which compose a group convolution with unitary transforms in feature space to learn a richer set of representations than group convolution alone. UGConvs generalize two disparate ideas in CNN architecture, channel shuffling (i.e. ShuffleNet) and block-circulant networks (i.e. CirCNN), and provide unifying insights that lead to a deeper understanding of each technique. We experimentally demonstrate that dense unitary transforms can outperform channel shuffling in DNN accuracy. On t he other hand, different dense transforms exhibit comparable accuracy performanc e. Based on these observations we propose HadaNet, a UGConv network using Hadama rd transforms. HadaNets achieve similar accuracy to circulant networks with lowe r computation complexity, and better accuracy than ShuffleNets with the same num ber of parameters and floating-point multiplies.
****

Semi-Supervised Learning With Graph Learning-Convolutional Networks
Bo Jiang, Ziyan Zhang, Doudou Lin, Jin Tang, Bin Luo; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11 313-11320
Graph Convolutional Neural Networks (graph CNNs) have been widely used for graph data representation and semi-supervised learning tasks. However, existing graph CNNs generally use a fixed graph which may not be optimal for semi-supervised l earning tasks. In this paper, we propose a novel Graph Learning-Convolutional Ne twork (GLCN) for graph data representation and semi-supervised learning. The aim of GLCN is to learn an optimal graph structure that best serves graph CNNs for semi-supervised learning by integrating both graph learning and graph convolutio n in a unified network architecture. The main advantage is that in GLCN both giv en labels and the estimated labels are incorporated and thus can provide useful 'weakly' supervised information to refine (or learn) the graph construction and also to facilitate the graph convolution operation for unknown label estimation. Experimental results on seven benchmarks demonstrate that GLCN significantly ou tperforms the state-of-the-art traditional fixed structure based graph CNNs.
****

Learning to Remember: A Synaptic Plasticity Driven Framework for Continual Learn ing
Oleksiy Ostapenko, Mihai Puscas, Tassilo Klein, Patrick Jahnichen, Moin Nabi ; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit ion (CVPR), 2019, pp. 11321-11329
Models trained in the context of continual learning (CL) should be able to learn from a stream of data over an undefined period of time. The main challenges her ein are: 1) maintaining old knowledge while simultaneously benefiting from it wh en learning new tasks, and 2) guaranteeing model scalability with a growing amou nt of data to learn from. In order to tackle these challenges, we introduce Dyna mic Generative Memory (DGM) - synaptic plasticity driven framework for continual learning. DGM relies on conditional generative adversarial networks with learna ble connection plasticity realized with neural masking. Specifically, we evaluat e two variants of neural masking: applied to (i) layer activations and (ii) to c onnection weights directly. Furthermore, we propose a dynamic network expansion mechanism that ensures sufficient model capacity to accommodate for continually incoming tasks. The amount of added capacity is determined dynamically from the

learned binary mask. We evaluate DGM in the continual class-incremental setup on visual classification tasks.

****

AIRD: Adversarial Learning Framework for Image Repurposing Detection

Ayush Jaiswal, Yue Wu, Wael AbdAlmageed, Iacopo Masi, Premkumar Natarajan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11330-11339

Image repurposing is a commonly used method for spreading misinformation on social media and online forums, which involves publishing untampered images with modified metadata to create rumors and further propaganda. While manual verification is possible, given vast amounts of verified knowledge available on the internet, the increasing prevalence and ease of this form of semantic manipulation call for the development of robust automatic ways of assessing the semantic integrity of multimedia data. In this paper, we present a novel method for image repurposing detection that is based on the real-world adversarial interplay between a bad actor who repurposes images with counterfeit metadata and a watchdog who verifies the semantic consistency between images and their accompanying metadata, where both players have access to a reference dataset of verified content, which they can use to achieve their goals. The proposed method exhibits state-of-the-art performance on location-identity, subject-identity and painting-artist verification, showing its efficacy across a diverse set of scenarios.

****

A Kernelized Manifold Mapping to Diminish the Effect of Adversarial Perturbations

Saeid Asgari Taghanaki, Kumar Abhishek, Shekoofeh Azizi, Ghassan Hamarneh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11340-11349

The linear and non-flexible nature of deep convolutional models makes them vulnerable to carefully crafted adversarial perturbations. To tackle this problem, we propose a non-linear radial basis convolutional feature mapping by learning a Mahalanobis-like distance function. Our method then maps the convolutional features onto a linearly well-separated manifold, which prevents small adversarial perturbations from forcing a sample to cross the decision boundary. We test the proposed method on three publicly available image classification and segmentation datasets namely, MNIST, ISBI ISIC 2017 skin lesion segmentation, and NIH Chest X-Ray-14. We evaluate the robustness of our method to different gradient (targeted and untargeted) and non-gradient based attacks and compare it to several non-gradient masking defense strategies. Our results demonstrate that the proposed method can increase the resilience of deep convolutional neural networks to adversarial perturbations without accuracy drop on clean data.

****

Trust Region Based Adversarial Attack on Neural Networks

Zhewei Yao, Amir Gholami, Peng Xu, Kurt Keutzer, Michael W. Mahoney; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11350-11359

Deep Neural Networks are quite vulnerable to adversarial perturbations. Current state-of-the-art adversarial attack methods typically require very time consuming hyper-parameter tuning, or require many iterations to solve an optimization based adversarial attack. To address this problem, we present a new family of trust region based adversarial attacks, with the goal of computing adversarial perturbations efficiently. We propose several attacks based on variants of the trust region optimization method. We test the proposed methods on Cifar-10 and ImageNet datasets using several different models including AlexNet, ResNet-50, VGG-16, and DenseNet-121 models. Our methods achieve comparable results with the Carlini-Wagner (CW) attack, but with significant speed up of up to 37x, for the VGG-16 model on a Titan Xp GPU. For the case of ResNet-50 on ImageNet, we can bring down its classification accuracy to less than 0.1% with at most 1.5% relative L_infinity (or L_2) perturbation requiring only 1.02 seconds as compared to 27.04 seconds for the CW attack. We have open sourced our method which can be accessed at [??].

****
PEPSI : Fast Image Inpainting With Parallel Decoding Network
Min-cheol Sagong, Yong-goo Shin, Seung-wook Kim, Seung Park, Sung-jea Ko; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11360-11368
Recently, a generative adversarial network (GAN)-based method employing the coarse-to-fine network with the contextual attention module (CAM) has shown outstanding results in image inpainting. However, this method requires numerous computational resources due to its two-stage process for feature encoding. To solve this problem, in this paper, we present a novel network structure, called PEPSI: parallel extended-decoder path for semantic inpainting. PEPSI can reduce the number of convolution operations by adopting a structure consisting of a single shared encoding network and a parallel decoding network with coarse and inpainting paths. The coarse path produces a preliminary inpainting result with which the encoding network is trained to predict features for the CAM. At the same time, the inpainting path creates a higher-quality inpainting result using refined features reconstructed by the CAM. PEPSI not only reduces the number of convolution operation almost by half as compared to the conventional coarse-to-fine networks but also exhibits superior performance to other models in terms of testing time and qualitative scores.
****
Model-Blind Video Denoising via Frame-To-Frame Training
Thibaud Ehret, Axel Davy, Jean-Michel Morel, Gabriele Facciolo, Pablo Arias; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11369-11378
Modeling the processing chain that has produced a video is a difficult reverse engineering task, even when the camera is available. This makes model based video processing a still more complex task. In this paper we propose a fully blind video denoising method, with two versions off-line and on-line. This is achieved by fine-tuning a pre-trained AWGN denoising network to the video with a novel frame-to-frame training strategy. Our denoiser can be used without knowledge of the origin of the video or burst and the post-processing steps applied from the camera sensor. The on-line process only requires a couple of frames before achieving visually pleasing results for a wide range of perturbations. It nonetheless reaches state-of-the-art performance for standard Gaussian noise, and can be used off-line with still better performance.
****
End-To-End Efficient Representation Learning via Cascading Combinatorial Optimization
Yeonwoo Jeong, Yoonsung Kim, Hyun Oh Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11379-11387
We develop hierarchically quantized efficient embedding representations for similarity-based search and show that this representation provides not only the state of the art performance on the search accuracy but also provides several orders of speed up during inference. The idea is to hierarchically quantize the representation so that the quantization granularity is greatly increased while maintaining the accuracy and keeping the computational complexity low. We also show that the problem of finding the optimal sparse compound hash code respecting the hierarchical structure can be optimized in polynomial time via minimum cost flow in an equivalent flow network. This allows us to train the method end-to-end in a mini-batch stochastic gradient descent setting. Our experiments on Cifar100 and ImageNet datasets show the state of the art search accuracy while providing several orders of magnitude search speedup respectively over exhaustive linear search over the dataset.
****
Sim-Real Joint Reinforcement Transfer for 3D Indoor Navigation
Fengda Zhu, Linchao Zhu, Yi Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11388-11397
There has been an increasing interest in 3D indoor navigation, where a robot in an environment moves to a target according to an instruction. To deploy a robot

for navigation in the physical world, lots of training data is required to learn an effective policy. It is quite labour intensive to obtain sufficient real environment data for training robots while synthetic data is much easier to construct by render-ing. Though it is promising to utilize the synthetic environments to facilitate navigation training in the real world, real environment are heterogeneous from synthetic environment in two aspects. First, the visual representation of the two environments have significant variances. Second, the houseplans of these two environments are quite different. There-fore two types of information,i.e. visual representation and policy behavior, need to be adapted in the reinforce mentmodel. The learning procedure of visual representation and that of policy behavior are presumably reciprocal. We pro-pose to jointly adapt visual representation and policy behavior to leverage the mutual impacts of environment and policy. Specifically, our method employs an adversarial feature adaptation model for visual representation transfer anda policy mimic strategy for policy behavior imitation. Experiment shows that our method outperforms the baseline by 19.47% without any additional human annotations.

****

ChamNet: Towards Efficient Network Design Through Platform-Aware Model Adaptation

Xiaoliang Dai, Peizhao Zhang, Bichen Wu, Hongxu Yin, Fei Sun, Yanghan Wang, Marat Dukhan, Yunqing Hu, Yiming Wu, Yangqing Jia, Peter Vajda, Matt Uyttendaele, Niraj K. Jha; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11398-11407

This paper proposes an efficient neural network (NN) architecture design methodology called Chameleon that honors given resource constraints. Instead of developing new building blocks or using computationally-intensive reinforcement learning algorithms, our approach leverages existing efficient network building blocks and focuses on exploiting hardware traits and adapting computation resources to fit target latency and/or energy constraints. We formulate platform-aware NN architecture search in an optimization framework and propose a novel algorithm to search for optimal architectures aided by efficient accuracy and resource (latency and/or energy) predictors. At the core of our algorithm lies an accuracy predictor built atop Gaussian Process with Bayesian optimization for iterative sampling. With a one-time building cost for the predictors, our algorithm produces state-of-the-art model architectures on different platforms under given constraints in just minutes. Our results show that adapting computation resources to building blocks is critical to model performance. Without the addition of any special features, our models achieve significant accuracy improvements relative to state-of-the-art handcrafted and automatically designed architectures. We achieve 73.8% and 75.3% top-1 accuracy on ImageNet at 20ms latency on a mobile CPU and DSP. At reduced latency, our models achieve up to 8.2% (4.8%) and 6.7% (9.3%) absolute top-1 accuracy improvements compared to MobileNetV2 and MnasNet, respectively, on a mobile CPU (DSP), and 2.7% (4.6%) and 5.6% (2.6%) accuracy gains over ResNet-101 and ResNet-152, respectively, on an Nvidia GPU (Intel CPU).

****

Regularizing Activation Distribution for Training Binarized Deep Networks

Ruizhou Ding, Ting-Wu Chin, Zeye Liu, Diana Marculescu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11408-11417

Binarized Neural Networks (BNNs) can significantly reduce the inference latency and energy consumption in resource-constrained devices due to their pure-logical computation and fewer memory accesses. However, training BNNs is difficult since the activation flow encounters degeneration, saturation, and gradient mismatch problems. Prior work alleviates these issues by increasing activation bits and adding floating-point scaling factors, thereby sacrificing BNN's energy efficiency. In this paper, we propose to use distribution loss to explicitly regularize the activation flow, and develop a framework to systematically formulate the loss. Our experiments show that the distribution loss can consistently improve the accuracy of BNNs without losing their energy benefits. Moreover, equipped with the proposed regularization, BNN training is shown to be robust to the selection

of hyper-parameters including optimizer and learning rate.
****
Robustness Verification of Classification Deep Neural Networks via Linear Programming
Wang Lin, Zhengfeng Yang, Xin Chen, Qingye Zhao, Xiangkun Li, Zhiming Liu, Jifeng He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11418-11427
There is a pressing need to verify robustness of classification deep neural networks (CDNNs) as they are embedded in many safety-critical applications. Existing robustness verification approaches rely on computing the over-approximation of the output set, and can hardly scale up to practical CDNNs, as the result of error accumulation accompanied with approximation. In this paper, we develop a novel method for robustness verification of CDNNs with sigmoid activation functions. It converts the robustness verification problem into an equivalent problem of inspecting the most suspected point in the input region which constitutes a nonlinear optimization problem. To make it amenable, by relaxing the nonlinear constraints into the linear inclusions, it is further refined as a linear programming problem. We conduct comparison experiments on a few CDNNs trained for classifying images in some state-of-the-art benchmarks, showing our advantages of precision and scalability that enable effective verification of practical CDNNs.
****
Additive Adversarial Learning for Unbiased Authentication
Jian Liang, Yuren Cao, Chenbin Zhang, Shiyu Chang, Kun Bai, Zenglin Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11428-11437
Authentication is a task aiming to confirm the truth between data instances and personal identities. Typical authentication applications include face recognition, person re-identification, authentication based on mobile devices and so on. The recently-emerging data-driven authentication process may encounter undesired biases, i.e., the models are often trained in one domain (e.g., for people wearing spring outfits) while required to apply in other domains (e.g., they change the clothes to summer outfits). To address this issue, we propose a novel two-stage method that disentangles the class/identity from domain-differences, and we consider multiple types of domain-difference. In the first stage, we learn disentangled representations by a one-versus-rest disentangle learning (OVRDL) mechanism. In the second stage, we improve the disentanglement by an additive adversarial learning (AAL) mechanism. Moreover, we discuss the necessity to avoid a learning dilemma due to disentangling causally related types of domain-difference. Comprehensive evaluation results demonstrate the effectiveness and superiority of the proposed method.
****
Simultaneously Optimizing Weight and Quantizer of Ternary Neural Network Using Truncated Gaussian Approximation
Zhezhi He, Deliang Fan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11438-11446
In the past years, Deep convolution neural network has achieved great success in many artificial intelligence applications. However, its enormous model size and massive computation cost have become the main obstacle for deployment of such powerful algorithm in the low power and resource-limited mobile systems. As the countermeasure to this problem, deep neural networks with ternarized weights (i.e. -1, 0, +1) have been widely explored to greatly reduce model size and computational cost, with limited accuracy degradation. In this work, we propose a novel ternarized neural network training method which simultaneously optimizes both weights and quantizer during training, differentiating from prior works. Instead of fixed and uniform weight ternarization, we are the first to incorporate the thresholds of weight ternarization into a closed-form representation using truncated Gaussian approximation, enabling simultaneous optimization of weights and quantizer through back-propagation training. With both of the first and last layer ternarized, the experiments on the ImageNet classification task show that our ternarized ResNet-18/34/50 only has 3.9/2.52/2.16% accuracy degradation in compari

son to the full-precision counterparts.
****
Adversarial Defense by Stratified Convolutional Sparse Coding
Bo Sun, Nian-Hsuan Tsai, Fangchen Liu, Ronald Yu, Hao Su; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11447-11456
We propose an adversarial defense method that achieves state-of-the-art performance among attack-agnostic adversarial defense methods while also maintaining robustness to input resolution, scale of adversarial perturbation, and scale of dataset size. Based on convolutional sparse coding, we construct a stratified low-dimensional quasi-natural image space that faithfully approximates the natural image space while also removing adversarial perturbations. We introduce a novel Sparse Transformation Layer (STL) in between the input image and the first layer of the neural network to efficiently project images into our quasi-natural image space. Our experiments show state-of-the-art performance of our method compared to other attack-agnostic adversarial defense methods in various adversarial settings.
****
Exploring Object Relation in Mean Teacher for Cross-Domain Detection
Qi Cai, Yingwei Pan, Chong-Wah Ngo, Xinmei Tian, Lingyu Duan, Ting Yao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11457-11466
Rendering synthetic data (e.g., 3D CAD-rendered images) to generate annotations for learning deep models in vision tasks has attracted increasing attention in recent years. However, simply applying the models learnt on synthetic images may lead to high generalization error on real images due to domain shift. To address this issue, recent progress in cross-domain recognition has featured the Mean Teacher, which directly simulates unsupervised domain adaptation as semi-supervised learning. The domain gap is thus naturally bridged with consistency regularization in a teacher-student scheme. In this work, we advance this Mean Teacher paradigm to be applicable for cross-domain detection. Specifically, we present Mean Teacher with Object Relations (MTOR) that novelly remolds Mean Teacher under the backbone of Faster R-CNN by integrating the object relations into the measure of consistency cost between teacher and student modules. Technically, MTOR firstly learns relational graphs that capture similarities between pairs of regions for teacher and student respectively. The whole architecture is then optimized with three consistency regularizations: 1) region-level consistency to align the region-level predictions between teacher and student, 2) inter-graph consistency for matching the graph structures between teacher and student, and 3) intra-graph consistency to enhance the similarity between regions of same class within the graph of student. Extensive experiments are conducted on the transfers across Cityscapes, Foggy Cityscapes, and SIM10k, and superior results are reported when comparing to state-of-the-art approaches. More remarkably, we obtain a new record of single model: 22.8% of mAP on Syn2Real detection dataset.
****
Hierarchical Disentanglement of Discriminative Latent Features for Zero-Shot Learning
Bin Tong, Chao Wang, Martin Klinkigt, Yoshiyuki Kobayashi, Yuuichi Nonaka; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11467-11476
Most studies in zero-shot learning model the relationship, in the form of a classifier or mapping, between features from images of seen classes and their attributes. Therefore, the degree of a model's generalization ability for recognizing unseen images is highly constrained by that of image features and attributes. In this paper, we discuss two questions about generalization that are seldom discussed. Are image features trained with samples of seen classes expressive enough to capture the discriminative information for both seen and unseen classes? Is the relationship learned from seen image features and attributes sufficiently generalized to recognize unseen classes. To answer these two questions, we propose a model to learn discriminative and generalizable representations from image fea

tures under an auto-encoder framework. The discriminative latent features are le
arned through a group-wise disentanglement over feature groups with a hierarchic
al structure. On popular benchmark data sets, a significant improvement over sta
te-of-the-art methods in tasks of typical and generalized zero-shot learning ver
ifies the generalization ability of latent features for recognizing unseen image
s.
****

R2GAN: Cross-Modal Recipe Retrieval With Generative Adversarial Network
Bin Zhu, Chong-Wah Ngo, Jingjing Chen, Yanbin Hao; Proceedings of the IEEE/CV
F Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11477-
11486
Representing procedure text such as recipe for crossmodal retrieval is inherentl
y a difficult problem, not mentioning to generate image from recipe for visualiz
ation. This paper studies a new version of GAN, named Recipe Retrieval Generativ
e Adversarial Network (R2GAN), to explore the feasibility of generating image fr
om procedure text for retrieval problem. The motivation of using GAN is twofold:
 learning compatible cross-modal features in an adversarial way, and explanation
 of search results by showing the images generated from recipes. The novelty of
R2GAN comes from architecture design, specifically a GAN with one generator and
dual discriminators is used, which makes the generation of image from recipe a f
easible idea. Furthermore, empowered by the generated images, a two-level rankin
g loss in both embedding and image spaces are considered. These add-ons not only
 result in excellent retrieval performance, but also generate close-to-realistic
 food images useful for explaining ranking of recipes. On recipe1M dataset, R2GA
N demonstrates high scalability to data size, outperforms all the existing appro
aches, and generates images intuitive for human to interpret the search results.
****

Rethinking Knowledge Graph Propagation for Zero-Shot Learning
Michael Kampffmeyer, Yinbo Chen, Xiaodan Liang, Hao Wang, Yujia Zhang, Eric
 P. Xing; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
Recognition (CVPR), 2019, pp. 11487-11496
Graph convolutional neural networks have recently shown great potential for the
task of zero-shot learning. These models are highly sample efficient as related
concepts in the graph structure share statistical strength allowing generalizati
on to new classes when faced with a lack of data. However, multi-layer architect
ures, which are required to propagate knowledge to distant nodes in the graph, d
ilute the knowledge by performing extensive Laplacian smoothing at each layer an
d thereby consequently decrease performance. In order to still enjoy the benefit
 brought by the graph structure while preventing dilution of knowledge from dist
ant nodes, we propose a Dense Graph Propagation (DGP) module with carefully desi
gned direct links among distant nodes. DGP allows us to exploit the hierarchical
 graph structure of the knowledge graph through additional connections. These co
nnections are added based on a node's relationship to its ancestors and descenda
nts. A weighting scheme is further used to weigh their contribution depending on
 the distance to the node to improve information propagation in the graph. Combi
ned with finetuning of the representations in a two-stage training approach our
method outperforms state-of-the-art zero-shot learning approaches.
****

Learning to Learn Image Classifiers With Visual Analogy
Linjun Zhou, Peng Cui, Shiqiang Yang, Wenwu Zhu, Qi Tian; Proceedings of the
 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp
. 11497-11506
Humans are far better learners who can learn a new concept very fast with only a
 few samples compared with machines. The plausible mystery making the difference
 is two fundamental learning mechanisms: learning to learn and learning by analo
gy. In this paper, we attempt to investigate a new human-like learning method by
 organically combining these two mechanisms. In particular, we study how to gene
ralize the classification parameters from previously learned concepts to a new c
oncept. we first propose a novel Visual Analogy Graph Embedded Regression (VAGER
) model to jointly learn a low-dimensional embedding space and a linear mapping

function from the embedding space to classification parameters for base classes. We then propose an out-of-sample embedding method to learn the embedding of a new class represented by a few samples through its visual analogy with base classes and derive the classification parameters for the new class. We conduct extensive experiments on ImageNet dataset and the results show that our method could consistently and significantly outperform state-of-the-art baselines.
****

Where's Wally Now? Deep Generative and Discriminative Embeddings for Novelty Detection
Philippe Burlina, Neil Joshi, I-Jeng Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11507-11516
We develop a framework for novelty detection (ND) methods relying on deep embeddings, either discriminative or generative, and also propose a novel framework for assessing their performance. While much progress was made recently in these approaches, it has been accompanied by certain limitations: most methods were tested on relatively simple problems (low resolution images / small number of classes) or involved non-public data; comparative performance has often proven inconclusive because of lacking statistical significance; and evaluation has generally been done on non-canonical problem sets of differing complexity, making apples-to-apples comparative performance evaluation difficult. This has led to a relative confusing state of affairs. We address these challenges via the following contributions: We make a proposal for a novel framework to measure the performance of novelty detection methods using a trade-space demonstrating performance (measured by ROCAUC) as a function of problem complexity. We also make several proposals to formally characterize problem complexity. We conduct experiments with problems of higher complexity (higher image resolution / number of classes). To this end we design several canonical datasets built from CIFAR-10 and ImageNet (IN-125) which we make available to perform future benchmarks for novelty detection as well as other related tasks including semantic zero/adaptive shot and unsupervised learning. Finally, we demonstrate, as one of the methods in our ND framework, a generative novelty detection method whose performance exceeds that of all recent best-in-class generative ND methods.
****

Weakly Supervised Image Classification Through Noise Regularization
Mengying Hu, Hu Han, Shiguang Shan, Xilin Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11517-11525
Weakly supervised learning is an essential problem in computer vision tasks, such as image classification, object recognition, etc., because it is expected to work in the scenarios where a large dataset with clean labels is not available. While there are a number of studies on weakly supervised image classification, they usually limited to either single-label or multi-label scenarios. In this work, we propose an effective approach for weakly supervised image classification utilizing massive noisy labeled data with only a small set of clean labels (e.g., 5%). The proposed approach consists of a clean net and a residual net, which aim to learn a mapping from feature space to clean label space and a residual mapping from feature space to the residual between clean labels and noisy labels, respectively, in a multi-task learning manner. Thus, the residual net works as a regularization term to improve the clean net training. We evaluate the proposed approach on two multi-label datasets (OpenImage and MS COCO2014) and a single-label dataset (Clothing1M). Experimental results show that the proposed approach outperforms the state-of-the-art methods, and generalizes well to both single-label and multi-label scenarios.
****

Data-Driven Neuron Allocation for Scale Aggregation Networks
Yi Li, Zhanghui Kuang, Yimin Chen, Wayne Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11526-11534
Successful visual recognition networks benefit from aggregating information spanning from a wide range of scales. Previous research has investigated information

fusion of connected layers or multiple branches in a block, seeking to strengthen the power of multi-scale representations. Despite their great successes, existing practices often allocate the neurons for each scale manually, and keep the same ratio in all aggregation blocks of an entire network, rendering suboptimal performance. In this paper, we propose to learn the neuron allocation for aggregating multi-scale information in different building blocks of a deep network. The most informative output neurons in each block are preserved while others are discarded, and thus neurons for multiple scales are competitively and adaptively allocated. Our scale aggregation network (ScaleNet) is constructed by repeating a scale aggregation (SA) block that concatenates feature maps at a wide range of scales. Feature maps for each scale are generated by a stack of downsampling, convolution and upsampling operations. The data-driven neuron allocation and SA block achieve strong representational power at the cost of considerably low computational complexity. The proposed ScaleNet, by replacing all 3x3 convolutions in ResNet with our SA blocks, achieves better performance than ResNet and its outstanding variants like ResNeXt and SE-ResNet, in the same computational complexity. On ImageNet classification, ScaleNets absolutely reduce the top-1 error rate of ResNets by 1.12 (101 layers) and 1.82 (50 layers). On COCO object detection, ScaleNets absolutely improve the mAP with backbone of ResNets by 3.6 and 4.6 on Faster-RCNN, respectively. Code and models are released on https://github.com/Eli-YiLi/ScaleNet.

****

Graphical Contrastive Losses for Scene Graph Parsing

Ji Zhang, Kevin J. Shih, Ahmed Elgammal, Andrew Tao, Bryan Catanzaro; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11535-11543

Most scene graph parsers use a two-stage pipeline to detect visual relationships: the first stage detects entities, and the second predicts the predicate for each entity pair using a softmax distribution. We find that such pipelines, trained with only a cross entropy loss over predicate classes, suffer from two common errors. The first, Entity Instance Confusion, occurs when the model confuses multiple instances of the same type of entity (e.g. multiple cups). The second, Proximal Relationship Ambiguity, arises when multiple subject-predicate-object triplets appear in close proximity with the same predicate, and the model struggles to infer the correct subject-object pairings (e.g. mis-pairing musicians and their instruments). We propose a set of contrastive loss formulations that specifically target these types of errors within the scene graph parsing problem, collectively termed the Graphical Contrastive Losses. These losses explicitly force the model to disambiguate related and unrelated instances through margin constraints specific to each type of confusion. We further construct a relationship detector, called RelDN, using the aforementioned pipeline to demonstrate the efficacy of our proposed losses. Our model outperforms the winning method of the OpenImages Relationship Detection Challenge by 4.7% (16.5% relatively) on the test set. We also show improved results over the best previous methods on the Visual Genome and Visual Relationship Detection datasets.

****

Deep Transfer Learning for Multiple Class Novelty Detection

Pramuditha Perera, Vishal M. Patel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11544-11552

We propose a transfer learning-based solution for the problem of multiple class novelty detection. In particular, we propose an end-to-end deep-learning based approach in which we investigate how the knowledge contained in an external, out-of-distributional dataset can be used to improve the performance of a deep network for visual novelty detection. Our solution differs from the standard deep classification networks on two accounts. First, we use a novel loss function, membership loss, in addition to the classical cross-entropy loss for training networks. Secondly, we use the knowledge from the external dataset more effectively to learn globally negative filters, filters that respond to generic objects outside the known class set. We show that thresholding the maximal activation of the proposed network can be used to identify novel objects effectively. Extensive

experiments on four publicly available novelty detection datasets show that the proposed method achieves significant improvements over the state-of-the-art methods.
****

QATM: Quality-Aware Template Matching for Deep Learning
Jiaxin Cheng, Yue Wu, Wael AbdAlmageed, Premkumar Natarajan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11553-11562
Finding a template in a search image is one of the core problems in many computer vision applications, such as template matching, image semantic alignment, image-to-GPS verification etc.. In this paper, we propose a novel quality-aware template matching method, which is not only used as a standalone template matching algorithm, but also a trainable layer that can be easily plugged in any deep neural network. Specifically, we assess the quality of a matching pair as its soft-ranking among all matching pairs, and thus different matching scenarios like 1-to-1, 1-to-many, and many-to-many will be all reflected to different values. Our extensive studies in the classic template matching problem and deep learning tasks demonstrate the effectiveness of QATM: it not only outperforms SOTA template matching methods when used alone, but also largely improves existing DNN solutions when used in DNN.
****

Retrieval-Augmented Convolutional Neural Networks Against Adversarial Examples
Jake Zhao (Junbo), Kyunghyun Cho; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11563-11571
We propose a retrieval-augmented convolutional network (RaCNN) and propose to train it with local mixup, a novel variant of the recently proposed mixup algorithm. The proposed hybrid architecture combining a convolutional network and an off-the-shelf retrieval engine was designed to mitigate the adverse effect of off-manifold adversarial examples, while the proposed local mixup addresses on-manifold ones by explicitly encouraging the classifier to locally behave linearly on the data manifold. Our evaluation of the proposed approach against seven readilyavailable adversarial attacks on three datasets-CIFAR-10, SVHN and ImageNet-demonstrate the improved robustness compared to a vanilla convolutional network, and comparable performance with the state-of-the-art reactive defense approaches.
****

Learning Cross-Modal Embeddings With Adversarial Networks for Cooking Recipes and Food Images
Hao Wang, Doyen Sahoo, Chenghao Liu, Ee-peng Lim, Steven C. H. Hoi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11572-11581
Food computing is playing an increasingly important role in human daily life, and has found tremendous applications in guiding human behavior towards smart food consumption and healthy lifestyle. An important task under the food-computing umbrella is retrieval, which is particularly helpful for health related applications, where we are interested in retrieving important information about food (e.g., ingredients, nutrition, etc.). In this paper, we investigate an open research task of cross-modal retrieval between cooking recipes and food images, and propose a novel framework Adversarial Cross-Modal Embedding (ACME) to resolve the cross-modal retrieval task in food domains. Specifically, the goal is to learn a common embedding feature space between the two modalities, in which our approach consists of several novel ideas: (i) learning by using a new triplet loss scheme together with an effective sampling strategy, (ii) imposing modality alignment using an adversarial learning strategy, and (iii) imposing cross-modal translation consistency such that the embedding of one modality is able to recover some important information of corresponding instances in the other modality. ACME achieves the state-of-the-art performance on the benchmark Recipe1M dataset, validating the efficacy of the proposed technique.
****

FastDraw: Addressing the Long Tail of Lane Detection by Adapting a Sequential Prediction Network

Jonah Philion; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11582-11591

The search for predictive models that generalize to the long tail of sensor inputs is the central difficulty when developing data-driven models for autonomous vehicles. In this paper, we use lane detection to study modeling and training techniques that yield better performance on real world test drives. On the modeling side, we introduce a novel fully convolutional model of lane detection that learns to decode lane structures instead of delegating structure inference to post-processing. In contrast to previous works, our convolutional decoder is able to represent an arbitrary number of lanes per image, preserves the polyline representation of lanes without reducing lanes to polynomials, and draws lanes iteratively without requiring the computational and temporal complexity of recurrent neural networks. Because our model includes an estimate of the joint distribution of neighboring pixels belonging to the same lane, our formulation includes a natural and computationally cheap definition of uncertainty. On the training side, we demonstrate a simple yet effective approach to adapt the model to new environments using unsupervised style transfer. By training FastDraw to make predictions of lane structure that are invariant to low-level stylistic differences between images, we achieve strong performance at test time in weather and lighting conditions that deviate substantially from those of the annotated datasets that are publicly available. We quantitatively evaluate our approach on the CVPR 2017 Tusimple lane marking challenge, difficult CULane datasets [29], and a small labeled dataset of our own and achieve competitive accuracy while running at 90 FPS.

****

Weakly Supervised Video Moment Retrieval From Text Queries

Niluthpol Chowdhury Mithun, Sujoy Paul, Amit K. Roy-Chowdhury; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11592-11601

There have been a few recent methods proposed in text to video moment retrieval using natural language queries, but requiring full supervision during training. However, acquiring a large number of training videos with temporal boundary annotations for each text description is extremely time-consuming and often not scalable. In order to cope with this issue, in this work, we introduce the problem of learning from weak labels for the task of text to video moment retrieval. The weak nature of the supervision is because, during training, we only have access to the video-text pairs rather than the temporal extent of the video to which different text descriptions relate. We propose a joint visual-semantic embedding based framework that learns the notion of relevant segments from video using only video-level sentence descriptions. Specifically, our main idea is to utilize latent alignment between video frames and sentence descriptions using Text-Guided Attention (TGA). TGA is then used during the test phase to retrieve relevant moments. Experiments on two benchmark datasets demonstrate that our method achieves comparable performance to state-of-the-art fully supervised approaches.

****

Content-Aware Multi-Level Guidance for Interactive Instance Segmentation

Soumajit Majumder, Angela Yao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11602-11611

In interactive instance segmentation, users give feedback to iteratively refine segmentation masks. The user-provided clicks are transformed into guidance maps which provide the network with necessary cues on the whereabouts of the object of interest. Guidance maps used in current systems are purely distance-based and are either too localized or non-informative. We propose a novel transformation of user clicks to generate content-aware guidance maps that leverage the hierarchical structural information present in an image. Using our guidance maps, even the most basic FCNs are able to outperform existing approaches that require state-of-the-art segmentation networks pre-trained on large scale segmentation datasets. We demonstrate the effectiveness of our proposed transformation strategy through comprehensive experimentation in which we significantly raise state-of-the-art on four standard interactive segmentation benchmarks.

****

Greedy Structure Learning of Hierarchical Compositional Models
Adam Kortylewski, Aleksander Wieczorek, Mario Wieser, Clemens Blumer, Sonali Parbhoo, Andreas Morel-Forster, Volker Roth, Thomas Vetter; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11612-11621
In this work, we consider the problem of learning a hierarchical generative model of an object from a set of images which show examples of the object in the presence of variable background clutter. Existing approaches to this problem are limited by making strong a-priori assumptions about the object's geometric structure and require seg- mented training data for learning. In this paper, we propose a novel framework for learning hierarchical compositional models (HCMs) which do not suffer from the mentioned limitations. We present a generalized formulation of HCMs and describe a greedy structure learning framework that consists of two phases: Bottom-up part learning and top-down model composition. Our framework integrates the foreground-background segmentation problem into the structure learning task via a background model. As a result, we can jointly optimize for the number of layers in the hierarchy, the number of parts per layer and a foreground- background segmentation based on class labels only. We show that the learned HCMs are semantically meaningful and achieve competitive results when compared to other generative object models at object classification on a standard transfer learning dataset.
****
Interactive Full Image Segmentation by Considering All Regions Jointly
Eirikur Agustsson, Jasper R. R. Uijlings, Vittorio Ferrari; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11622-11631
We address interactive full image annotation, where the goal is to accurately segment all object and stuff regions in an image. We propose an interactive, scribble-based annotation framework which operates on the whole image to produce segmentations for all regions. This enables sharing scribble corrections across regions, and allows the annotator to focus on the largest errors made by the machine across the whole image. To realize this, we adapt Mask-RCNN [22] into a fast interactive segmentation framework and introduce an instance-aware loss measured at the pixel-level in the full image canvas, which lets predictions for nearby regions properly compete for space. Finally, we compare to interactive single object segmentation on the COCO panoptic dataset [11, 27, 34]. We demonstrate that our interactive full image segmentation approach leads to a 5% IoU gain, reaching 90% IoU at a budget of four extreme clicks and four corrective scribbles per region
****
Learning Active Contour Models for Medical Image Segmentation
Xu Chen, Bryan M. Williams, Srinivasa R. Vallabhaneni, Gabriela Czanner, Rachel Williams, Yalin Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11632-11640
Image segmentation is an important step in medical image processing and has been widely studied and developed for refinement of clinical analysis and applications. New models based on deep learning have improved results but are restricted to pixel-wise fitting of the segmentation map. Our aim was to tackle this limitation by developing a new model based on deep learning which takes into account the area inside as well as outside the region of interest as well as the size of boundaries during learning. Specifically, we propose a new loss function which incorporates area and size information and integrates this into a dense deep learning model. We evaluated our approach on a dataset of more than 2,000 cardiac MRI scans. Our results show that the proposed loss function outperforms other mainstream loss function Cross-entropy on two common segmentation networks. Our loss function is robust while using different hyperparameter lambda.
****
Customizable Architecture Search for Semantic Segmentation
Yiheng Zhang, Zhaofan Qiu, Jingen Liu, Ting Yao, Dong Liu, Tao Mei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR

), 2019, pp. 11641-11650
In this paper, we propose a Customizable Architecture Search (CAS) approach to a utomatically generate a network architecture for semantic image segmentation. Th e generated network consists of a sequence of stacked computation cells. A compu tation cell is represented as a directed acyclic graph, in which each node is a hidden representation (i.e., feature map) and each edge is associated with an op eration (e.g., convolution and pooling), which transforms data to a new layer. D uring the training, the CAS algorithm explores the search space for an optimized computation cell to build a network. The cells of the same type share one archi tecture but with different weights. In real applications, however, an optimizati on may need to be conducted under some constraints such as GPU time and model si ze. To this end, a cost corresponding to the constraint will be assigned to each operation. When an operation is selected during the search, its associated cost will be added to the objective. As a result, our CAS is able to search an optim ized architecture with customized constraints. The approach has been thoroughly evaluated on Cityscapes and CamVid datasets, and demonstrates superior performan ce over several state-of-the-art techniques. More remarkably, our CAS achieves 7 2.3% mIoU on the Cityscapes dataset with speed of 108 FPS on an Nvidia TitanXp G PU.
****

Local Features and Visual Words Emerge in Activations
Oriane Simeoni, Yannis Avrithis, Ondrej Chum; Proceedings of the IEEE/CVF Conf erence on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11651-11660
We propose a novel method of deep spatial matching (DSM) for image retrieval. In itial ranking is based on image descriptors extracted from convolutional neural network activations by global pooling, as in recent state-of-the-art work. Howev er, the same sparse 3D activation tensor is also approximated by a collection of local features. These local features are then robustly matched to approximate t he optimal alignment of the tensors. This happens without any network modificati on, additional layers or training. No local feature detection happens on the ori ginal image. No local feature descriptors and no visual vocabulary are needed th roughout the whole process. We experimentally show that the proposed method ach ieves the state-of-the-art performance on standard benchmarks across different n etwork architectures and different global pooling methods. The highest gain in p erformance is achieved when diffusion on the nearest-neighbor graph of global de scriptors is initiated from spatially verified images.
****

Hyperspectral Image Super-Resolution With Optimized RGB Guidance
Ying Fu, Tao Zhang, Yinqiang Zheng, Debing Zhang, Hua Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11661-11670
To overcome the limitations of existing hyperspectral cameras on spatial/tempo ral resolution, fusing a low resolution hyperspectral image (HSI) with a high resolution RGB (or multispectral) image into a high resolution HSI has been pr evalent. Previous methods for this fusion task usually employ hand-crafted pri ors to model the underlying structure of the latent high resolution HSI, and t he effect of the camera spectral response (CSR) of the RGB camera on super-res olution accuracy has rarely been investigated. In this paper, we first present a simple and efficient convolutional neural network (CNN) based method for HS I super-resolution in an unsupervised way, without any prior training. Later, we append a CSR optimization layer onto the HSI super-resolution network, eith er to automatically select the best CSR in a given CSR dataset, or to design t he optimal CSR under some physical restrictions. Experimental results show our method outperforms the state-of-the-arts, and the CSR optimization can furthe r boost the accuracy of HSI super-resolution.
****

Adaptive Confidence Smoothing for Generalized Zero-Shot Learning
Yuval Atzmon, Gal Chechik; Proceedings of the IEEE/CVF Conference on Computer V ision and Pattern Recognition (CVPR), 2019, pp. 11671-11680
Generalized zero-shot learning (GZSL) is the problem of learning a classifier wh

ere some classes have samples and others are learned from side information, like semantic attributes or text description, in a zero-shot learning fashion (ZSL). Training a single model that operates in these two regimes simultaneously is challenging. Here we describe a probabilistic approach that breaks the model into three modular components, and then combines them in a consistent way. Specifically, our model consists of three classifiers: A "gating" model that makes soft decisions if a sample is from a "seen" class, and two experts: a ZSL expert, and an expert model for seen classes. We address two main difficulties in this approach: How to provide an accurate estimate of the gating probability without any training samples for unseen classes; and how to use expert predictions when it observes samples outside of its domain. The key insight to our approach is to pass information between the three models to improve each one's accuracy, while maintaining the modular structure. We test our approach, adaptive confidence smoothing (COSMO), on four standard GZSL benchmark datasets and find that it largely outperforms state-of-the-art GZSL models. COSMO is also the first model that closes the gap and surpasses the performance of generative models for GZSL, even-though it is a light-weight model that is much easier to train and tune.
****

PMS-Net: Robust Haze Removal Based on Patch Map for Single Images
Wei-Ting Chen, Jian-Jiun Ding, Sy-Yen Kuo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11681-11689
In this paper, we proposed a novel haze removal algorithm based on a new feature called the patch map. Conventional patch-based haze removal algorithms (e.g. the Dark Channel prior) usually performs dehazing with a fixed patch size. However, it may produce several problems in recovered results such as oversaturation and color distortion. Therefore, in this paper, we designed an adaptive and automatic patch size selection model called the Patch Map Selection Network (PMS-Net) to select the patch size corresponding to each pixel. This network is designed based on the convolutional neural network (CNN), which can generate the patch map from the image to image. Experimental results on both synthesized and real-world hazy images show that, with the combination of the proposed PMS-Net, the performance in haze removal is much better than that of other state-of-the-art algorithms and we can address the problems caused by the fixed patch size.
****

Deep Spherical Quantization for Image Search
Sepehr Eghbali, Ladan Tahvildari; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11690-11699
Hashing methods, which encode high-dimensional images with compact discrete codes, have been widely applied to enhance large-scale image retrieval. In this paper, we put forward Deep Spherical Quantization (DSQ), a novel method to make deep convolutional neural networks generate supervised and compact binary codes for efficient image search. Our approach simultaneously learns a mapping that transforms the input images into a low-dimensional discriminative space, and quantizes the transformed data points using multi-codebook quantization. To eliminate the negative effect of norm variance on codebook learning, we force the network to L_2 normalize the extracted features and then quantize the resulting vectors using a new supervised quantization technique specifically designed for points lying on a unit hypersphere. Furthermore, we introduce an easy-to-implement extension of our quantization technique that enforces sparsity on the codebooks. Extensive experiments demonstrate that DSQ and its sparse variant can generate semantically separable compact binary codes outperforming many state-of-the-art image retrieval methods on three benchmarks.
****

Large-Scale Interactive Object Segmentation With Human Annotators
Rodrigo Benenson, Stefan Popov, Vittorio Ferrari; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11700-11709
Manually annotating object segmentation masks is very time consuming. Interactive object segmentation methods offer a more efficient alternative where a human annotator and a machine segmentation model collaborate. In this paper we make sev

eral contributions to interactive segmentation: (1) we systematically explore in simulation the design space of deep interactive segmentation models and report new insights and caveats; (2) we execute a large-scale annotation campaign with real human annotators, producing masks for 2.5M instances on the OpenImages data set. We released this data publicly, forming the largest existing dataset for in stance segmentation. Moreover, by re-annotating part of the COCO dataset, we sho w that we can produce instance masks 3x faster than traditional polygon drawing tools while also providing better quality. (3) We present a technique for automa tically estimating the quality of the produced masks which exploits indirect sig nals from the annotation process.
****

A Poisson-Gaussian Denoising Dataset With Real Fluorescence Microscopy Images
Yide Zhang, Yinhao Zhu, Evan Nichols, Qingfei Wang, Siyuan Zhang, Cody Smit h, Scott Howard; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11710-11718
Fluorescence microscopy has enabled a dramatic development in modern biology. Du e to its inherently weak signal, fluorescence microscopy is not only much noisie r than photography, but also presented with Poisson-Gaussian noise where Poisson noise, or shot noise, is the dominating noise source. To get clean fluorescence microscopy images, it is highly desirable to have effective denoising algorithm s and datasets that are specifically designed to denoise fluorescence microscopy images. While such algorithms exist, no such datasets are available. In this pa per, we fill this gap by constructing a dataset - the Fluorescence Microscopy De noising (FMD) dataset - that is dedicated to Poisson-Gaussian denoising. The dat aset consists of 12,000 real fluorescence microscopy images obtained with commer cial confocal, two-photon, and wide-field microscopes and representative biologi cal samples such as cells, zebrafish, and mouse brain tissues. We use image aver aging to effectively obtain ground truth images and 60,000 noisy images with dif ferent noise levels. We use this dataset to benchmark 10 representative denoisin g algorithms and find that deep learning methods have the best performance. To o ur knowledge, this is the first real microscopy image dataset for Poisson-Gaussi an denoising purposes and it could be an important tool for high-quality, real-t ime denoising applications in biomedical research.
****

Task Agnostic Meta-Learning for Few-Shot Learning
Muhammad Abdullah Jamal, Guo-Jun Qi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11719-11727
Meta-learning approaches have been proposed to tackle the few-shot learning prob lem. Typically, a meta-learner is trained on a variety of tasks in the hopes of being generalizable to new tasks. However, the generalizability on new tasks of a meta-learner could be fragile when it is over-trained on existing tasks durin g meta-training phase. In other words, the initial model of a meta-learner could be too biased towards existing tasks to adapt to new tasks, especially when onl y very few examples are available to update the model. To avoid a biased meta-le arner and improve its generalizability, we propose a novel paradigm of Task-Agno stic Meta-Learning (TAML) algorithms. Specifically, we present an entropy-based approach that meta-learns an unbiased initial model with the largest uncertainty over the output labels by preventing it from over-performing in classification tasks. Alternatively, a more general inequality-minimization TAML is presented f or more ubiquitous scenarios by directly minimizing the inequality of initial lo sses beyond the classification tasks wherever a suitable loss can be defined. Ex periments on benchmarked datasets demonstrate that the proposed approaches outpe rform compared meta-learning algorithms in both few-shot classification and rein forcement learning tasks.
****

Progressive Ensemble Networks for Zero-Shot Recognition
Meng Ye, Yuhong Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11728-11736
Despite the advancement of supervised image recognition algorithms, their depend ence on the availability of labeled data and the rapid expansion of image catego

ries raise the significant challenge of zero-shot learning. Zero-shot learning (ZSL) aims to transfer knowledge from labeled classes into unlabeled classes to r educe human labeling effort. In this paper, we propose a novel progressive ensem ble network model with multiple projected label embeddings to address zero-shot image recognition. The ensemble network is built by learning multiple image clas sification functions with a shared feature extraction network but different labe l embedding representations, which enhance the diversity of the classifiers and facilitate information transfer to unlabeled classes. A progressive training fra mework is then deployed to gradually label the most confident images in each unl abeled class with predicted pseudo-labels and update the ensemble network with t he training data augmented by the pseudo-labels. The proposed model performs tra ining on both labeled and unlabeled data. It can naturally bridge the domain shi ft problem in visual appearances and be extended to the generalized zero-shot le arning scenario. We conduct experiments on multiple ZSL datasets and the empiric al results demonstrate the efficacy of the proposed model.

****

Direct Object Recognition Without Line-Of-Sight Using Optical Coherence

Xin Lei, Liangyu He, Yixuan Tan, Ken Xingze Wang, Xinggang Wang, Yihan Du, Shanhui Fan, Zongfu Yu; Proceedings of the IEEE/CVF Conference on Computer Vis ion and Pattern Recognition (CVPR), 2019, pp. 11737-11746

Visual object recognition under situations in which the direct line-of-sight is blocked, such as when it is occluded around the corner, is of practical importan ce in a wide range of applications. With coherent illumination, the light scatte red from diffusive walls forms speckle patterns that contain information of the hidden object. It is possible to realize non-line-of-sight (NLOS) recognition wi th these speckle patterns. We introduce a novel approach based on speckle patter n recognition with deep neural network, which is simpler and more robust than ot her NLOS recognition methods. Simulations and experiments are performed to verif y the feasibility and performance of this approach.

****

Atlas of Digital Pathology: A Generalized Hierarchical Histological Tissue Type-Annotated Database for Deep Learning

Mahdi S. Hosseini, Lyndon Chan, Gabriel Tse, Michael Tang, Jun Deng, Sajad Norouzi, Corwyn Rowsell, Konstantinos N. Plataniotis, Savvas Damaskinos; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2019, pp. 11747-11756

In recent years, computer vision techniques have made large advances in image re cognition and been applied to aid radiological diagnosis. Computational patholog y aims to develop similar tools for aiding pathologists in diagnosing digitized histopathological slides, which would improve diagnostic accuracy and productivi ty amidst increasing workloads. However, there is a lack of publicly-available d atabases of (1) localized patch-level images annotated with (2) a large range of Histological Tissue Type (HTT). As a result, computational pathology research i s constrained to diagnosing specific diseases or classifying tissues from specif ic organs, and cannot be readily generalized to handle unexpected diseases and o rgans. In this paper, we propose a new digital pathology database, the "Atlas o f Digital Pathology" (or ADP), which comprises of 17,668 patch images extracted from 100 slides annotated with up to 57 hierarchical HTTs. Our data is generaliz ed to different tissue types across different organs and aims to provide trainin g data for supervised multi-label learning of patch-level HTT in a digitized who le slide image. We demonstrate the quality of our image labels through pathologi st consultation and by training three state-of-the-art neural networks on tissue type classification. Quantitative results support the visually consistency of o ur data and we demonstrate a tissue type-based visual attention aid as a sample tool that could be developed from our database.

****

Perturbation Analysis of the 8-Point Algorithm: A Case Study for Wide FoV Camera s

Thiago L. T. da Silveira, Claudio R. Jung; Proceedings of the IEEE/CVF Conferen ce on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11757-11766

This paper presents a perturbation analysis for the estimate of epipolar matrices using the 8-Point Algorithm (8-PA). Our approach explores existing bounds for singular subspaces and relates them to the 8-PA, without assuming any kind of error distribution for the matched features. In particular, if we use unit vectors as homogeneous image coordinates, we show that having a wide spatial distribution of matched features in both views tends to generate lower error bounds for the epipolar matrix error. Our experimental validation indicates that the bounds and the effective errors tend to decrease as the camera Field of View (FoV) increases, and that using the 8-PA for spherical images (that present 360degx180deg FoV) leads to accurate essential matrices. As an additional contribution, we present bounds for the direction of the translation vector extracted from the essential matrix based on singular subspace analysis.

****

Robustness of 3D Deep Learning in an Adversarial Setting

Matthew Wicker, Marta Kwiatkowska; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11767-11775

Understanding the spatial arrangement and nature of real-world objects is of paramount importance to many complex engineering tasks, including autonomous navigation. Deep learning has revolutionized state-of-the-art performance for tasks in 3D environments; however, relatively little is known about the robustness of these approaches in an adversarial setting. The lack of comprehensive analysis makes it difficult to justify deployment of 3D deep learning models in real-world, safety-critical applications. In this work, we develop an algorithm for analysis of pointwise robustness of neural networks that operate on 3D data. We show that current approaches presented for understanding the resilience of state-of-the-art models vastly overestimate their robustness. We then use our algorithm to evaluate an array of state-of-the-art models in order to demonstrate their vulnerability to occlusion attacks. We show that, in the worst case, these networks can be reduced to 0% classification accuracy after the occlusion of at most 6.5% of the occupied input space.

****

SceneCode: Monocular Dense Semantic Reconstruction Using Learned Encoded Scene Representations

Shuaifeng Zhi, Michael Bloesch, Stefan Leutenegger, Andrew J. Davison; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11776-11785

Systems which incrementally create 3D semantic maps from image sequences must store and update representations of both geometry and semantic entities. However, while there has been much work on the correct formulation for geometrical estimation, state-of-the-art systems usually rely on simple semantic representations which store and update independent label estimates for each surface element (depth pixels, surfels, or voxels). Spatial correlation is discarded, and fused label maps are incoherent and noisy. We introduce a new compact and optimisable semantic representation by training a variational auto-encoder that is conditioned on a colour image. Using this learned latent space, we can tackle semantic label fusion by jointly optimising the low-dimenional codes associated with each of a set of overlapping images, producing consistent fused label maps which preserve spatial correlation. We also show how this approach can be used within a monocular keyframe based semantic mapping system where a similar code approach is used for geometry. The probabilistic formulation allows a flexible formulation where we can jointly estimate motion, geometry and semantics in a unified optimisation.

****

StereoDRNet: Dilated Residual StereoNet

Rohan Chabra, Julian Straub, Christopher Sweeney, Richard Newcombe, Henry Fuchs; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11786-11795

We propose a system that uses a convolution neural network (CNN) to estimate depth from a stereo pair followed by volumetric fusion of the predicted depth maps to produce a 3D reconstruction of a scene. Our proposed depth refinement archite

cture, predicts view-consistent disparity and occlusion maps that helps the fusi on system to produce geometrically consistent reconstructions. We utilize 3D dil ated convolutions in our proposed cost filtering network that yields better filt ering while almost halving the computational cost in comparison to state of the art cost filtering architectures. For feature extraction we use the Vortex Pooli ng architecture. The proposed method achieves state of the art results in KITTI 2012, KITTI 2015 and ETH 3D stereo benchmarks. Finally, we demonstrate that our system is able to produce high fidelity 3D scene reconstructions that outperform s the state of the art stereo system.
****

The Alignment of the Spheres: Globally-Optimal Spherical Mixture Alignment for C amera Pose Estimation
Dylan Campbell, Lars Petersson, Laurent Kneip, Hongdong Li, Stephen Gould; P roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11796-11806
Determining the position and orientation of a calibrated camera from a single im age with respect to a 3D model is an essential task for many applications. When 2D-3D correspondences can be obtained reliably, perspective-n-point solvers can be used to recover the camera pose. However, without the pose it is non-trivial to find cross-modality correspondences between 2D images and 3D models, particul arly when the latter only contains geometric information. Consequently, the prob lem becomes one of estimating pose and correspondences jointly. Since outliers a nd local optima are so prevalent, robust objective functions and global search s trategies are desirable. Hence, we cast the problem as a 2D-3D mixture model ali gnment task and propose the first globally-optimal solution to this formulation under the robust L2 distance between mixture distributions. We derive novel boun ds on this objective function and employ branch-and-bound to search the 6D space of camera poses, guaranteeing global optimality without requiring a pose estima te. To accelerate convergence, we integrate local optimization, implement GPU bo und computations, and provide an intuitive way to incorporate side information s uch as semantic labels. The algorithm is evaluated on challenging synthetic and real datasets, outperforming existing approaches and reliably converging to the global optimum.
****

Learning Joint Reconstruction of Hands and Manipulated Objects
Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black , Ivan Laptev, Cordelia Schmid; Proceedings of the IEEE/CVF Conference on Comp uter Vision and Pattern Recognition (CVPR), 2019, pp. 11807-11816
Estimating hand-object manipulations is essential for in- terpreting and imitati ng human actions. Previous work has made significant progress towards reconstruc tion of hand poses and object shapes in isolation. Yet, reconstructing hands and objects during manipulation is a more challeng- ing task due to significant occ lusions of both the hand and object. While presenting challenges, manipulations may also simplify the problem since the physics of contact re- stricts the space of valid hand-object configurations. For example, during manipulation, the hand and object should be in contact but not interpenetrate. In this work, we regu- larize the joint reconstruction of hands and objects with ma- nipulation constra ints. We present an end-to-end learnable model that exploits a novel contact los s that favors phys- ically plausible hand-object constellations. Our approach im proves grasp quality metrics over baselines, using RGB images as input. To train and evaluate the model, we also propose a new large-scale synthetic dataset, Ob Man, with hand-object manipulations. We demonstrate the transfer- ability of ObM an-trained models to real data.
****

Deep Single Image Camera Calibration With Radial Distortion
Manuel Lopez, Roger Mari, Pau Gargallo, Yubin Kuang, Javier Gonzalez-Jimenez , Gloria Haro; Proceedings of the IEEE/CVF Conference on Computer Vision and Pa ttern Recognition (CVPR), 2019, pp. 11817-11825
Single image calibration is the problem of predicting the camera parameters from one image. This problem is of importance when dealing with images collected in

uncontrolled conditions by non-calibrated cameras, such as crowd-sourced applica
tions. In this work we propose a method to predict extrinsic (tilt and roll) and
 intrinsic (focal length and radial distortion) parameters from a single image.
We propose a parameterization for radial distortion that is better suited for le
arning than directly predicting the distortion parameters. Moreover, predicting
additional heterogeneous variables exacerbates the problem of loss balancing. We
 propose a new loss function based on point projections to avoid having to balan
ce heterogeneous loss terms. Our method is, to our knowledge, the first to joint
ly estimate the tilt, roll, focal length, and radial distortion parameters from
a single image. We thoroughly analyze the performance of the proposed method and
 the impact of the improvements and compare with previous approaches for single
image radial distortion correction.
****

CAM-Convs: Camera-Aware Multi-Scale Convolutions for Single-View Depth
Jose M. Facil, Benjamin Ummenhofer, Huizhong Zhou, Luis Montesano, Thomas Br
ox, Javier Civera; Proceedings of the IEEE/CVF Conference on Computer Vision an
d Pattern Recognition (CVPR), 2019, pp. 11826-11835
Single-view depth estimation suffers from the problem that a network trained on
images from one camera does not generalize to images taken with a different came
ra model. Thus, changing the camera model requires collecting an entirely new tr
aining dataset. In this work, we propose a new type of convolution that can take
 the camera parameters into account, thus allowing neural networks to learn cali
bration-aware patterns. Experiments confirm that this improves the generalizatio
n capabilities of depth prediction networks considerably, and clearly outperform
s the state of the art when the train and test images are acquired with differen
t cameras.
****

Translate-to-Recognize Networks for RGB-D Scene Recognition
Dapeng Du, Limin Wang, Huiling Wang, Kai Zhao, Gangshan Wu; Proceedings of t
he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019,
pp. 11836-11845
Cross-modal transfer is helpful to enhance modality-specific discriminative powe
r for scene recognition. To this end, this paper presents a unified framework to
 integrate the tasks of cross-modal translation and modality-specific recognitio
n, termed as Translate-to-Recognize Network TRecgNet. Specifically, both transla
tion and recognition tasks share the same encoder network, which allows to expli
citly regularize the training of recognition task with the help of translation,
and thus improve its final generalization ability. For translation task, we plac
e a decoder module on top of the encoder network and it is optimized with a new
layer-wise semantic loss, while for recognition task, we use a linear classifier
 based on the feature embedding from encoder and its training is guided by the s
tandard cross-entropy loss. In addition, our TRecgNet allows to exploit large nu
mbers of unlabeled RGB-D data to train the translation task and thus improve the
 representation power of encoder network. Empirically, we verify that this new s
emi-supervised setting is able to further enhance the performance of recognition
 network. We perform experiments on two RGB-D scene recognition benchmarks: NYU
Depth v2 and SUN RGB-D, demonstrating that TRecgNet achieves superior performanc
e to the existing state-of-the-art methods, especially for recognition solely ba
sed on a single modality.
****

Re-Identification Supervised Texture Generation
Jian Wang, Yunshan Zhong, Yachun Li, Chi Zhang, Yichen Wei; Proceedings of t
he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019,
pp. 11846-11856
The estimation of 3D human body pose and shape from a single image has been exte
nsively studied in recent years. However, the texture generation problem has not
 been fully discussed. In this paper, we propose an end-to-end learning strategy
 to generate textures of human bodies under the supervision of person re-identif
ication. We render the synthetic images with textures extracted from the inputs
and maximize the similarity between the rendered and input images by using the r

e-identification network as the perceptual metrics. Experiment results on pedest rian images show that our model can generate the texture from a single image and demonstrate that our textures are of higher quality than those generated by oth er available methods. Furthermore, we extend the application scope to other cate gories and explore the possible utilization of our generated textures.
****

Action4D: Online Action Recognition in the Crowd and Clutter
Quanzeng You, Hao Jiang; Proceedings of the IEEE/CVF Conference on Computer Vis ion and Pattern Recognition (CVPR), 2019, pp. 11857-11866
Recognizing every person's action in a crowded and cluttered environment is a ch allenging task in computer vision. We propose to tackle this challenging problem using a holistic 4D "scan" of a cluttered scene to include every detail about t he people and environment. This leads to a new problem, i.e., recognizing multip le people's actions in the cluttered 4D representation. At the first step, we pr opose a new method to track people in 4D, which can reliably detect and follow e ach person in real time. Then, we build a new deep neural network, the Action4DN et, to recognize the action of each tracked person. Such a model gives reliable and accurate results in the real-world settings. We also design an adaptive 3D c onvolution layer and a novel discriminative temporal feature learning objective to further improve the performance of our model. Our method is invariant to came ra view angles, resistant to clutter and able to handle crowd. The experimental results show that the proposed method is fast, reliable and accurate. Our method paves the way to action recognition in the real-world applications and is ready to be deployed to enable smart homes, smart factories and smart stores.
****

Monocular 3D Object Detection Leveraging Accurate Proposals and Shape Reconstruc tion
Jason Ku, Alex D. Pon, Steven L. Waslander; Proceedings of the IEEE/CVF Confer ence on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11867-11876
We present MonoPSR, a monocular 3D object detection method that leverages propos als and shape reconstruction. First, using the fundamental relations of a pinhol e camera model, detections from a mature 2D object detector are used to generate a 3D proposal per object in a scene. The 3D location of these proposals prove t o be quite accurate, which greatly reduces the difficulty of regressing the fina l 3D bounding box detection. Simultaneously, a point cloud is predicted in an ob ject centered coordinate system to learn local scale and shape information. How ever, the key challenge is how to exploit shape information to guide 3D localiza tion. As such, we devise aggregate losses, including a novel projection alignmen t loss, to jointly optimize these tasks in the neural network to improve 3D loca lization accuracy. We validate our method on the KITTI benchmark where we set ne w state-of-the-art results among published monocular methods, including the hard er pedestrian and cyclist classes, while maintaining efficient run-time.
****

Attribute-Aware Face Aging With Wavelet-Based Generative Adversarial Networks
Yunfan Liu, Qi Li, Zhenan Sun; Proceedings of the IEEE/CVF Conference on Compu ter Vision and Pattern Recognition (CVPR), 2019, pp. 11877-11886
Since it is difficult to collect face images of the same subject over a long ran ge of age span, most existing face aging methods resort to unpaired datasets to learn age mappings. However, the matching ambiguity between young and aged face images inherent to unpaired training data may lead to unnatural changes of facia l attributes during the aging process, which could not be solved by only enforci ng identity consistency like most existing studies do. In this paper, we propose an attribute-aware face aging model with wavelet based Generative Adversarial N etworks (GANs) to address the above issues. To be specific, we embed facial attr ibute vectors into both the generator and discriminator of the model to encourag e each synthesized elderly face image to be faithful to the attribute of its cor responding input. In addition, a wavelet packet transform (WPT) module is incorp orated to improve the visual fidelity of generated images by capturing age-relat ed texture details at multiple scales in the frequency space. Qualitative result s demonstrate the ability of our model in synthesizing visually plausible face i

mages, and extensive quantitative evaluation results show that the proposed meth od achieves state-of-the-art performance on existing datasets.
****

Noise-Tolerant Paradigm for Training Face Recognition CNNs
Wei Hu, Yangyu Huang, Fan Zhang, Ruirui Li; Proceedings of the IEEE/CVF Confe rence on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11887-11896
Benefit from large-scale training datasets, deep Convolutional Neural Networks(C NNs) have achieved impressive results in face recognition(FR). However, tremendo us scale of datasets inevitably lead to noisy data, which obviously reduce the p erformance of the trained CNN models. Kicking out wrong labels from large-scale FR datasets is still very expensive, although some cleaning approaches are propo sed. According to the analysis of the whole process of training CNN models super vised by angular margin based loss(AM-Loss) functions, we find that the distrib ution of training samples implicitly reflects their probability of being clean. Thus, we propose a novel training paradigm that employs the idea of weighting sa mples based on the above probability. Without any prior knowledge of noise, we c an train high performance CNN models with largescale FR datasets. Experiments de monstrate the effectiveness of our training paradigm. The codes are available at https://github.com/huangyangyu/NoiseFace.
****

Low-Rank Laplacian-Uniform Mixed Model for Robust Face Recognition
Jiayu Dong, Huicheng Zheng, Lina Lian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11897-11906
Sparse representation based methods have successfully put forward a general fram ework for robust face recognition through linear reconstruction and sparsity con straints. However, residual modeling in existing works is not yet robust enough when dealing with dense noise. In this paper, we aim at recognizing identities f rom faces with varying levels of noises of various forms such as occlusion, pixe l corruption, or disguise, and take improving the fitting ability of the error m odel as the key to addressing this problem. To fully capture the characteristics of different noises, we propose a mixed model combining robust sparsity constra int and low-rank constraint, which can deal with random errors and structured er rors simultaneously. For random noises such as pixel corruption, we adopt a Lapl acian-uniform mixed function for fitting the error distribution. For structured errors like continuous occlusion or disguise, we utilize robust nuclear norm to constrain the rank of the error matrix. An effective iterative reweighted algori thm is then developed to solve the proposed model. Comprehensive experiments wer e conducted on several benchmark databases for robust face recognition, and the overall results demonstrate that our model is most robust against various kinds of noises, when compared with state-of-the-art methods.
****

Generalizing Eye Tracking With Bayesian Adversarial Learning
Kang Wang, Rui Zhao, Hui Su, Qiang Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11907-11916
Existing appearance-based gaze estimation approaches with CNN have poor generali zation performance. By systematically studying this issue, we identify three maj or factors: 1) appearance variations; 2) head pose variations and 3) over-fittin g issue with point estimation. To improve the generalization performance, we pro pose to incorporate adversarial learning and Bayesian inference into a unified f ramework. In particular, we first add an adversarial component into traditional CNN-based gaze estimator so that we can learn features that are gaze-responsive but can generalize to appearance and pose variations. Next, we extend the point-estimation based deterministic model to a Bayesian framework so that gaze estima tion can be performed using all parameters instead of only one set of parameters . Besides improved performance on several benchmark datasets, the proposed metho d also enables online adaptation of the model to new subjects/environments, demo nstrating the potential usage for practical real-time eye tracking applications.
****

Local Relationship Learning With Person-Specific Shape Regularization for Facial Action Unit Detection

Xuesong Niu, Hu Han, Songfan Yang, Yan Huang, Shiguang Shan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11917-11926

Encoding individual facial expressions via action units (AUs) coded by the Facial Action Coding System (FACS) has been found to be an effective approach in resolving the ambiguity issue among different expressions. While a number of methods have been proposed for AU detection, robust AU detection in the wild remains a challenging problem because of the diverse baseline AU intensities across individual subjects, and the weakness of appearance signal of AUs. To resolve these issues, in this work, we propose a novel AU detection method by utilizing local information and the relationship of individual local face regions. Through such a local relationship learning, we expect to utilize rich local information to improve the AU detection robustness against the potential perceptual inconsistency of individual local regions. In addition, considering the diversity in the baseline AU intensities of individual subjects, we further regularize local relationship learning via person-specific face shape information, i.e., reducing the influence of person-specific shape information, and obtaining more AU discriminative features. The proposed approach outperforms the state-of-the-art methods on two widely used AU detection datasets in the public domain (BP4D and DISFA).

****

Point-To-Pose Voting Based Hand Pose Estimation Using Residual Permutation Equivariant Layer

Shile Li, Dongheui Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11927-11936

Recently, 3D input data based hand pose estimation methods have shown state-of-the-art performance, because 3D data capture more spatial information than the depth image. Whereas 3D voxel-based methods need a large amount of memory, PointNet based methods need tedious preprocessing steps such as K-nearest neighbour search for each point. In this paper, we present a novel deep learning hand pose estimation method for an unordered point cloud. Our method takes 1024 3D points as input and does not require additional information. We use Permutation Equivariant Layer (PEL) as the basic element, where a residual network version of PEL is proposed for the hand pose estimation task. Furthermore, we propose a voting-based scheme to merge information from individual points to the final pose output. In addition to the pose estimation task, the voting-based scheme can also provide point cloud segmentation result without ground-truth for segmentation. We evaluate our method on both NYU dataset and the Hands2017Challenge dataset, where our method outperforms recent state-of-theart methods.

****

Improving Few-Shot User-Specific Gaze Adaptation via Gaze Redirection Synthesis

Yu Yu, Gang Liu, Jean-Marc Odobez; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11937-11946

As an indicator of human attention gaze is a subtle behavioral cue which can be exploited in many applications. However, inferring 3D gaze direction is challenging even for deep neural networks given the lack of large amount of data (groundtruthing gaze is expensive and existing datasets use different setups) and the inherent presence of gaze biases due to person-specific difference. In this work, we address the problem of person-specific gaze model adaptation from only a few reference training samples. The main and novel idea is to improve gaze adaptation by generating additional training samples through the synthesis of gaze-redirected eye images from existing reference samples. In doing so, our contributions are threefold:(i) we design our gaze redirection framework from synthetic data, allowing us to benefit from aligned training sample pairs to predict accurate inverse mapping fields; (ii) we proposed a self-supervised approach for domain adaptation; (iii) we exploit the gaze redirection to improve the performance of person-specific gaze estimation. Extensive experiments on two public datasets demonstrate the validity of our gaze retargeting and gaze estimation framework.

****

AdaptiveFace: Adaptive Margin and Sampling for Face Recognition

Hao Liu, Xiangyu Zhu, Zhen Lei, Stan Z. Li; Proceedings of the IEEE/CVF Confe

rence on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11947-11956
Training large-scale unbalanced data is the central topic in face recognition. I n the past two years, face recognition has achieved remarkable improvements due to the introduction of margin based Softmax loss. However, these methods have an implicit assumption that all the classes possess sufficient samples to describe its distribution, so that a manually set margin is enough to equally squeeze ea ch intra-class variations. However, real face datasets are highly unbalanced, wh ich means the classes have tremendously different numbers of samples. In this pa per, we argue that the margin should be adapted to different classes. We propose the Adaptive Margin Softmax to adjust the margins for different classes adaptiv ely. In addition to the unbalance challenge, face data always consists of large- scale classes and samples. Smartly selecting valuable classes and samples to par ticipate in the training makes the training more effective and efficient. To thi s end, we also make the sampling process adaptive in two folds: Firstly, we prop ose the Hard Prototype Mining to adaptively select a small number of hard classe s to participate in classification. Secondly, for data sampling, we introduce th e Adaptive Data Sampling to find valuable samples for training adaptively. We co mbine these three parts together as AdaptiveFace. Extensive analysis and experim ents on LFW, LFW BLUFR and MegaFace show that our method performs better than st ate-of-the-art methods using the same network architecture and training dataset. Code is available at https://github.com/haoliu1994/AdaptiveFace.
****

Disentangled Representation Learning for 3D Face Shape
Zi-Hang Jiang, Qianyi Wu, Keyu Chen, Juyong Zhang; Proceedings of the IEEE/CV F Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11957- 11966
In this paper, we present a novel strategy to design disentangled 3D face shape representation. Specifically, a given 3D face shape is decomposed into identity part and expression part, which are both encoded and decoded in a nonlinear way. To solve this problem, we propose an attribute decomposition framework for 3D f ace mesh. To better represent face shapes which are usually nonlinear deformed b etween each other, the face shapes are represented by a vertex based deformation representation rather than Euclidean coordinates. The experimental results demo nstrate that our method has better performance than existing methods on decompos ing the identity and expression parts. Moreover, more natural expression transfe r results can be achieved with our method than existing methods.
****

LBS Autoencoder: Self-Supervised Fitting of Articulated Meshes to Point Clouds
Chun-Liang Li, Tomas Simon, Jason Saragih, Barnabas Poczos, Yaser Sheikh; Pr oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11967-11976
We present LBS-AE; a self-supervised autoencoding algorithm for fitting articula ted mesh models to point clouds. As input, we take a sequence of point clouds to be registered as well as an artist-rigged mesh, i.e. a template mesh equipped w ith a linear-blend skinning (LBS) deformation space parameterized by a skeleton hierarchy. As output, we learn an LBS-based autoencoder that produces registered meshes from the input point clouds. To bridge the gap between the artist-define d geometry and the captured point clouds, our autoencoder models pose-dependent deviations from the template geometry. During training, instead of us- ing expli cit correspondences, such as key points or pose supervision, our method leverage s LBS deformations to boot- strap the learning process. To avoid poor local mini ma from erroneous point-to-point correspondences, we utilize a structured Chamfe r distance based on part-segmentations, which are learned concurrently using sel f-supervision. We demonstrate qualitative results on real captured hands, and re port quantitative evaluations on the FAUST benchmark for body registration. Our method achieves performance that is superior to other unsupervised approaches an d com- parable to methods using supervised examples.
****

PifPaf: Composite Fields for Human Pose Estimation
Sven Kreiss, Lorenzo Bertoni, Alexandre Alahi; Proceedings of the IEEE/CVF Con

ference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11977-11986
We propose a new bottom-up method for multi-person 2D human pose estimation that is particularly well suited for urban mobility such as self-driving cars and delivery robots. The new method, PifPaf, uses a Part Intensity Field (PIF) to localize body parts and a Part Association Field (PAF) to associate body parts with each other to form full human poses. Our method outperforms previous methods at low resolution and in crowded, cluttered and occluded scenes thanks to (i) our new composite field PAF encoding fine-grained information and (ii) the choice of Laplace loss for regressions which incorporates a notion of uncertainty. Our architecture is based on a fully convolutional, single-shot, box-free design. We perform on par with the existing state-of-the-art bottom-up method on the standard COCO keypoint task and produce state-of-the-art results on a modified COCO keypoint task for the transportation domain.
****

TACNet: Transition-Aware Context Network for Spatio-Temporal Action Detection
Lin Song, Shiwei Zhang, Gang Yu, Hongbin Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11987-11995
Current state-of-the-art approaches for spatio-temporal action detection have achieved impressive results but remain unsatisfactory for temporal extent detection. The main reason comes from that, there are some ambiguous states similar to the real actions which may be treated as target actions even by a well trained network. In this paper, we define these ambiguous samples as "transitional states", and propose a Transition-Aware Context Network (TACNet) to distinguish transitional states. The proposed TACNet includes two main components, i.e., temporal context detector and transition-aware classifier. The temporal context detector can extract long-term context information with constant time complexity by constructing a recurrent network. The transition-aware classifier can further distinguish transitional states by classifying action and transitional states simultaneously. Therefore, the proposed TACNet can substantially improve the performance of spatio-temporal action detection. We extensively evaluate the proposed TACNet on UCF101-24 and J-HMDB datasets. The experimental results demonstrate that TACNet obtains competitive performance on JHMDB and significantly outperforms the state-of-the-art methods on the untrimmed UCF101 24 in terms of both frame-mAP and video-mAP.
****

Learning Regularity in Skeleton Trajectories for Anomaly Detection in Videos
Romero Morais, Vuong Le, Truyen Tran, Budhaditya Saha, Moussa Mansour, Svetha Venkatesh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11996-12004
Appearance features have been widely used in video anomaly detection even though they contain complex entangled factors. We propose a new method to model the normal patterns of human movements in surveillance video for anomaly detection using dynamic skeleton features. We decompose the skeletal movements into two sub-components: global body movement and local body posture. We model the dynamics and interaction of the coupled features in our novel Message-Passing Encoder-Decoder Recurrent Network. We observed that the decoupled features collaboratively interact in our spatio-temporal model to accurately identify human-related irregular events from surveillance video sequences. Compared to traditional appearance-based models, our method achieves superior outlier detection performance. Our model also offers "open-box" examination and decision explanation made possible by the semantically understandable features and a network architecture supporting interpretability.
****

Local Temporal Bilinear Pooling for Fine-Grained Action Parsing
Yan Zhang, Siyu Tang, Krikamol Muandet, Christian Jarvers, Heiko Neumann; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12005-12015
Fine-grained temporal action parsing is important in many applications, such as daily activity understanding, human motion analysis, surgical robotics and others requiring subtle and precise operations over a long-term period. In this paper

we propose a novel bilinear pooling operation, which is used in intermediate la
yers of a temporal convolutional encoder-decoder net. In contrast to previous wo
rk, our proposed bilinear pooling is learnable and hence can capture more comple
x local statistics than the conventional counterpart. In addition, we introduce
exact lower-dimension representations of our bilinear forms, so that the dimensi
onality is reduced without suffering from information loss nor requiring extra c
omputation. We perform extensive experiments to quantitatively analyze our model
and show the superior performances to other state-of-the-art pooling work on va
rious datasets.
****

Improving Action Localization by Progressive Cross-Stream Cooperation
Rui Su, Wanli Ouyang, Luping Zhou, Dong Xu; Proceedings of the IEEE/CVF Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12016-12025
Spatio-temporal action localization consists of three levels of tasks: spatial l
ocalization, action classification, and temporal segmentation. In this work, we
propose a new Progressive Cross-stream Cooperation (PCSC) framework to iterative
improve action localization results and generate better bounding boxes for one
stream (i.e., Flow/RGB) by leveraging both region proposals and features from a
nother stream (i.e., RGB/Flow) in an iterative fashion. Specifically, we first g
enerate a larger set of region proposals by combining the latest region proposal
s from both streams, from which we can readily obtain a larger set of labelled t
raining samples to help learn better action detection models. Second, we also pr
opose a new message passing approach to pass information from one stream to anot
her stream in order to learn better representations, which also leads to better
action detection models. As a result, our iterative framework progressively impr
oves action localization results at the frame level. To improve action localizat
ion results at the video level, we additionally propose a new strategy to train
class-specific actionness detectors for better temporal segmentation, which can
be readily learnt by using the training samples around temporal boundaries. Comp
rehensive experiments on two benchmark datasets UCF-101-24 and J-HMDB demonstrat
e the effectiveness of our newly proposed approaches for spatio-temporal action
localization in realistic scenarios.
****

Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recog
nition
Lei Shi, Yifan Zhang, Jian Cheng, Hanqing Lu; Proceedings of the IEEE/CVF Con
ference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12026-12035
In skeleton-based action recognition, graph convolutional networks (GCNs), which
model the human body skeletons as spatiotemporal graphs, have achieved remarkab
le performance. However, in existing GCN-based methods, the topology of the grap
h is set manually, and it is fixed over all layers and input samples. This may n
ot be optimal for the hierarchical GCN and diverse samples in action recognition
tasks. In addition, the second-order information (the lengths and directions of
bones) of the skeleton data, which is naturally more informative and discrimina
tive for action recognition, is rarely investigated in existing methods. In this
work, we propose a novel two-stream adaptive graph convolutional network (2s-AG
CN) for skeleton-based action recognition. The topology of the graph in our mode
l can be either uniformly or individually learned by the BP algorithm in an end-
to-end manner. This data-driven method increases the flexibility of the model fo
r graph construction and brings more generality to adapt to various data samples
. Moreover, a two-stream framework is proposed to model both the first-order and
the second-order information simultaneously, which shows notable improvement fo
r the recognition accuracy. Extensive experiments on the two large-scale dataset
s, NTU-RGBD and Kinetics-Skeleton, demonstrate that the performance of our model
exceeds the state-of-the-art with a significant margin.
****

A Neural Network Based on SPD Manifold Learning for Skeleton-Based Hand Gesture
Recognition
Xuan Son Nguyen, Luc Brun, Olivier Lezoray, Sebastien Bougleux; Proceedings o
f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 201

9, pp. 12036-12045
This paper proposes a new neural network based on SPD manifold learning for skeleton-based hand gesture recognition. Given the stream of hand's joint positions, our approach combines two aggregation processes on respectively spatial and temporal domains. The pipeline of our network architecture consists in three main stages. The first stage is based on a convolutional layer to increase the discriminative power of learned features. The second stage relies on different architectures for spatial and temporal Gaussian aggregation of joint features. The third stage learns a final SPD matrix from skeletal data. A new type of layer is proposed for the third stage, based on a variant of stochastic gradient descent on Stiefel manifolds. The proposed network is validated on two challenging datasets and shows state-of-the-art accuracies on both datasets.
****

Large-Scale Weakly-Supervised Pre-Training for Video Action Recognition
Deepti Ghadiyaram, Du Tran, Dhruv Mahajan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12046-12055
Current fully-supervised video datasets consist of only a few hundred thousand videos and fewer than a thousand domain-specific labels. This hinders the progress towards advanced video architectures. This paper presents an in-depth study of using large volumes of web videos for pre-training video models for the task of action recognition. Our primary empirical finding is that pre-training at a very large scale (over 65 million videos), despite on noisy social-media videos and hashtags, substantially improves the state-of-the-art on three challenging public action recognition datasets. Further, we examine three questions in the construction of weakly-supervised video action datasets. First, given that actions involve interactions with objects, how should one construct a verb-object pre-training label space to benefit transfer learning the most? Second, frame-based models perform quite well on action recognition; is pre-training for good image features sufficient or is pre-training for spatio-temporal features valuable for optimal transfer learning? Finally, actions are generally less well-localized in long videos vs. short videos; since action labels are provided at a video level, how should one choose video clips for best performance, given some fixed budget of number or minutes of videos?
****

Learning Spatio-Temporal Representation With Local and Global Diffusion
Zhaofan Qiu, Ting Yao, Chong-Wah Ngo, Xinmei Tian, Tao Mei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12056-12065
Convolutional Neural Networks (CNN) have been regarded as a powerful class of models for visual recognition problems. Nevertheless, the convolutional filters in these networks are local operations while ignoring the large-range dependency. Such drawback becomes even worse particularly for video recognition, since video is an information-intensive media with complex temporal variations. In this paper, we present a novel framework to boost the spatio-temporal representation learning by Local and Global Diffusion (LGD). Specifically, we construct a novel neural network architecture that learns the local and global representations in parallel. The architecture is composed of LGD blocks, where each block updates local and global features by modeling the diffusions between these two representations. Diffusions effectively interact two aspects of information, i.e., localized and holistic, for more powerful way of representation learning. Furthermore, a kernelized classifier is introduced to combine the representations from two aspects for video recognition. Our LGD networks achieve clear improvements on the large-scale Kinetics-400 and Kinetics-600 video classification datasets against the best competitors by 3.5% and 0.7%. We further examine the generalization of both the global and local representations produced by our pre-trained LGD networks on four different benchmarks for video action recognition and spatio-temporal action detection tasks. Superior performances over several state-of-the-art techniques on these benchmarks are reported.
****

Unsupervised Learning of Action Classes With Continuous Temporal Embedding

Anna Kukleva, Hilde Kuehne, Fadime Sener, Jurgen Gall; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12066-12074

The task of temporally detecting and segmenting actions in untrimmed videos has seen an increased attention recently. One problem in this context arises from the need to define and label action boundaries to create annotations for training which is very time and cost intensive. To address this issue, we propose an unsupervised approach for learning action classes from untrimmed video sequences. To this end, we use a continuous temporal embedding of framewise features to benefit from the sequential nature of activities. Based on the latent space created by the embedding, we identify clusters of temporal segments across all videos that correspond to semantic meaningful action classes. The approach is evaluated on three challenging datasets, namely the Breakfast dataset, YouTube Instructions, and the 50Salads dataset. While previous works assumed that the videos contain the same high level activity, we furthermore show that the proposed approach can also be applied to a more general setting where the content of the videos is unknown.

****

Double Nuclear Norm Based Low Rank Representation on Grassmann Manifolds for Clustering

Xinglin Piao, Yongli Hu, Junbin Gao, Yanfeng Sun, Baocai Yin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12075-12084

Unsupervised clustering for high-dimension data (such as imageset or video) is a hard issue in data processing and data mining area since these data always lie on a manifold (such as Grassmann manifold). Inspired of Low Rank representation theory, researchers proposed a series of effective clustering methods for high-dimension data with non-linear metric. However, most of these methods adopt the traditional single nuclear norm as the relaxation of the rank function, which would lead to suboptimal solution deviated from the original one. In this paper, we propose a new low rank model for high-dimension data clustering task on Grassmann manifold based on the Double Nuclear norm which is used to better approximate the rank minimization of matrix. Further, to consider the inner geometry or structure of data space, we integrated the adaptive Laplacian regularization to construct the local relationship of data samples. The proposed models have been assessed on several public datasets for imageset clustering. The experimental results show that the proposed models outperform the state-of-the-art clustering ones.

****

SR-LSTM: State Refinement for LSTM Towards Pedestrian Trajectory Prediction

Pu Zhang, Wanli Ouyang, Pengfei Zhang, Jianru Xue, Nanning Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12085-12094

In crowd scenarios, reliable trajectory prediction of pedestrians requires insightful understanding of their social behaviors. These behaviors have been well investigated by plenty of studies, while it is hard to be fully expressed by hand-craft rules. Recent studies based on LSTM networks have shown great ability to learn social behaviors. However, many of these methods rely on previous neighboring hidden states but ignore the important current intention of the neighbors. In order to address this issue, we propose a data-driven state refinement module for LSTM network (SR-LSTM), which activates the utilization of the current intention of neighbors, and jointly and iteratively refines the current states of all participants in the crowd through a message passing mechanism. To effectively extract the social effect of neighbors, we further introduce a social-aware information selection mechanism consisting of an element-wise motion gate and a pedestrian-wise attention to select useful message from neighboring pedestrians. Experimental results on two public datasets, i.e. ETH and UCY, demonstrate the effectiveness of our proposed SR-LSTM and we achieve state-of-the-art results.

****

Unsupervised Deep Epipolar Flow for Stationary or Dynamic Scenes

Yiran Zhong, Pan Ji, Jianyuan Wang, Yuchao Dai, Hongdong Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12095-12104

Unsupervised deep learning for optical flow computation has achieved promising results. Most existing deep-net based methods rely on image brightness consistency and local smoothness constraint to train the networks. Their performance degrades at regions where repetitive textures or occlusions occur. In this paper, we propose Deep Epipolar Flow, an unsupervised optical flow method which incorporates global geometric constraints into network learning. In particular, we investigate multiple ways of enforcing the epipolar constraint in flow estimation. To alleviate a "chicken-and-egg" type of problem encountered in dynamic scenes where multiple motions may be present, we propose a low-rank constraint as well as a union-of-subspaces constraint for training. Experimental results on various benchmarking datasets show that our method achieves competitive performance compared with supervised methods and outperforms state-of-the-art unsupervised deep-learning methods.

****

An Efficient Schmidt-EKF for 3D Visual-Inertial SLAM

Patrick Geneva, James Maley, Guoquan Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12105-12115

It holds great implications for practical applications to enable centimeter-accuracy positioning for mobile and wearable sensor systems. In this paper, we propose a novel, high-precision, efficient visual-inertial (VI)-SLAM algorithm, termed Schmidt-EKF VI-SLAM (SEVIS), which optimally fuses IMU measurements and monocular images in a tightly-coupled manner to provide 3D motion tracking with bounded error. In particular, we adapt the Schmidt Kalman filter formulation to selectively include informative features in the state vector while treating them as nuisance parameters (or Schmidt states) once they become matured. This change in modeling allows for significant computational savings by no longer needing to constantly update the Schmidt states (or their covariance), while still allowing the EKF to correctly account for their cross-correlations with the active states. As a result, we achieve linear computational complexity in terms of map size, instead of quadratic as in the standard SLAM systems. In order to fully exploit the map information to bound navigation drifts, we advocate efficient keyframe-aided 2D-to-2D feature matching to find reliable correspondences between current 2D visual measurements and 3D map features. The proposed SEVIS is extensively validated in both simulations and experiments.

****

A Neural Temporal Model for Human Motion Prediction

Anand Gopalakrishnan, Ankur Mali, Dan Kifer, Lee Giles, Alexander G. Ororbia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12116-12125

We propose novel neural temporal models for predicting and synthesizing human motion, achieving state-of-the-art in modeling long-term motion trajectories while being competitive with prior work in short-term prediction and requiring significantly less computation. Key aspects of our proposed system include: 1) a novel, two-level processing architecture that aids in generating planned trajectories, 2) a simple set of easily computable features that integrate derivative information, and 3) a novel multi-objective loss function that helps the model to slowly progress from simple next-step prediction to the harder task of multi-step, closed-loop prediction. Our results demonstrate that these innovations improve the modeling of long-term motion trajectories. Finally, we propose a novel metric, called Normalized Power Spectrum Similarity (NPSS), to evaluate the long-term predictive ability of motion synthesis models, complementing the popular mean-squared error (MSE) measure of Euler joint angles over time. We conduct a user study to determine if the proposed NPSS correlates with human evaluation of long-term motion more strongly than MSE and find that it indeed does. We release code and additional results (visualizations) for this paper at: https://github.com/cr7anand/neural_temporal_models

****

Multi-Agent Tensor Fusion for Contextual Trajectory Prediction
Tianyang Zhao, Yifei Xu, Mathew Monfort, Wongun Choi, Chris Baker, Yibiao Zhao, Yizhou Wang, Ying Nian Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12126-12134
Accurate prediction of others' trajectories is essential for autonomous driving. Trajectory prediction is challenging because it requires reasoning about agents' past movements, social interactions among varying numbers and kinds of agents, constraints from the scene context, and the stochasticity of human behavior. Our approach models these interactions and constraints jointly within a novel Multi-Agent Tensor Fusion (MATF) network. Specifically, the model encodes multiple agents' past trajectories and the scene context into a Multi-Agent Tensor, then applies convolutional fusion to capture multiagent interactions while retaining the spatial structure of agents and the scene context. The model decodes recurrently to multiple agents' future trajectories, using adversarial loss to learn stochastic predictions. Experiments on both highway driving and pedestrian crowd datasets show that the model achieves state-of-the-art prediction accuracy.
****

Coordinate-Based Texture Inpainting for Pose-Guided Human Image Generation
Artur Grigorev, Artem Sevastopolsky, Alexander Vakhitov, Victor Lempitsky; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12135-12144
We present a new deep learning approach to pose-guided resynthesis of human photographs. At the heart of the new approach is the estimation of the complete body surface texture based on a single photograph. Since the input photograph always observes only a part of the surface, we suggest a new inpainting method that completes the texture of the human body. Rather than working directly with colors of texture elements, the inpainting network estimates an appropriate source location in the input image for each element of the body surface. This correspondence field between the input image and the texture is then further warped into the target image coordinate frame based on the desired pose, effectively establishing the correspondence between the source and the target view even when the pose change is drastic. The final convolutional network then uses the established correspondence and all other available information to synthesize the output image. A fully-convolutional architecture with deformable skip connections guided by the estimated correspondence field is used. We show state-of-the-art result for pose-guided image synthesis. Additionally, we demonstrate the performance of our system for garment transfer and pose-guided face resynthesis.
****

On Stabilizing Generative Adversarial Training With Noise
Simon Jenni, Paolo Favaro; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12145-12153
We present a novel method and analysis to train generative adversarial networks (GAN) in a stable manner. As shown in recent analysis, training is often undermined by the probability distribution of the data being zero on neighborhoods of the data space. We notice that the distributions of real and generated data should match even when they undergo the same filtering. Therefore, to address the limited support problem we propose to train GANs by using different filtered versions of the real and generated data distributions. In this way, filtering does not prevent the exact matching of the data distribution, while helping training by extending the support of both distributions. As filtering we consider adding samples from an arbitrary distribution to the data, which corresponds to a convolution of the data distribution with the arbitrary one. We also propose to learn the generation of these samples so as to challenge the discriminator in the adversarial training. We show that our approach results in a stable and well-behaved training of even the original minimax GAN formulation. Moreover, our technique can be incorporated in most modern GAN formulations and leads to a consistent improvement on several common datasets.
****

Self-Supervised GANs via Auxiliary Rotation Loss
Ting Chen, Xiaohua Zhai, Marvin Ritter, Mario Lucic, Neil Houlsby; Proceedin

gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12154-12163
Conditional GANs are at the forefront of natural image synthesis. The main drawback of such models is the necessity for labeled data. In this work we exploit two popular unsupervised learning techniques, adversarial training and self-supervision, and take a step towards bridging the gap between conditional and unconditional GANs. In particular, we allow the networks to collaborate on the task of representation learning, while being adversarial with respect to the classic GAN game. The role of self-supervision is to encourage the discriminator to learn meaningful feature representations which are not forgotten during training. We test empirically both the quality of the learned image representations, and the quality of the synthesized images. Under the same conditions, the self-supervised GAN attains a similar performance to state-of-the-art conditional counterparts. Finally, we show that this approach to fully unsupervised learning can be scaled to attain an FID of 23.4 on unconditional ImageNet generation.
****

Texture Mixer: A Network for Controllable Synthesis and Interpolation of Texture
Ning Yu, Connelly Barnes, Eli Shechtman, Sohrab Amirghodsi, Michal Lukac; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12164-12173
This paper addresses the problem of interpolating visual textures. We formulate this problem by requiring (1) by-example controllability and (2) realistic and smooth interpolation among an arbitrary number of texture samples. To solve it we propose a neural network trained simultaneously on a reconstruction task and a generation task, which can project texture examples onto a latent space where they can be linearly interpolated and projected back onto the image domain, thus ensuring both intuitive control and realistic results. We show our method outperforms a number of baselines according to a comprehensive suite of metrics as well as a user study. We further show several applications based on our technique, which include texture brush, texture dissolve, and animal hybridization.
****

Object-Driven Text-To-Image Synthesis via Adversarial Training
Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, Jianfeng Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12174-12182
In this paper, we propose Object-driven Attentive Generative Adversarial Newtorks (Obj-GANs) that allow attention-driven, multi-stage refinement for synthesizing complex images from text descriptions. With a novel object-driven attentive generative network, the Obj-GAN can synthesize salient objects by paying attention to their most relevant words in the text descriptions and their pre-generated class label. In addition, a novel object-wise discriminator based on the Fast R-CNN model is proposed to provide rich object-wise discrimination signals on whether the synthesized object matches the text description and the pre-generated class label. The proposed Obj-GAN significantly outperforms the previous state of the art in various metrics on the large-scale MS-COCO benchmark, increasing the inception score by 27% and decreasing the FID score by 11%. A thorough comparison between the classic grid attention and the new object-driven attention is provided through analyzing their mechanisms and visualizing their attention layers, showing insights of how the proposed model generates complex scenes in high quality.
****

Zoom-In-To-Check: Boosting Video Interpolation via Instance-Level Discrimination
Liangzhe Yuan, Yibo Chen, Hantian Liu, Tao Kong, Jianbo Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12183-12191
We propose a light-weight video frame interpolation algorithm. Our key innovation is an instance-level supervision that allows information to be learned from the high-resolution version of similar objects. Our experiment shows that the proposed method can generate state-of-the-art results across different datasets, with fractional computation resources (time and memory) of competing methods. Gi

ven two image frames, a cascade network creates an intermediate frame with 1) a flow-warping module that computes coarse bi-directional optical flow and creates an interpolated image via flow-based warping, followed by 2) an image synthesis module to make fine-scale corrections. In the learning stage, object detection proposals are generated on the interpolated image. Lower resolution objects are zoomed into, and the learning algorithms using an adversarial loss trained on high-resolution objects to guide the system towards the instance-level refinement corrects details of object shape and boundaries.
****

Disentangling Latent Space for VAE by Label Relevant/Irrelevant Dimensions
Zhilin Zheng, Li Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12192-12201
VAE requires the standard Gaussian distribution as a prior in the latent space. Since all codes tend to follow the same prior, it often suffers the so-called "posterior collapse". To avoid this, this paper introduces the class specific distribution for the latent code. But different from CVAE, we present a method for disentangling the latent space into the label relevant and irrelevant dimensions, zs and zu, for a single input. We apply two separated encoders to map the input into zs and zu respectively, and then give the concatenated code to the decoder to reconstruct the input. The label irrelevant code zu represent the common characteristics of all inputs, hence they are constrained by the standard Gaussian, and their encoder is trained in amortized variational inference way, like VAE. While zs is assumed to follow the Gaussian mixture distribution in which each component corresponds to a particular class. The parameters for the Gaussian components in zs encoder are optimized by the label supervision in a global stochastic way. In theory, we show that our method is actually equivalent to adding a KL divergence term on the joint distribution of zs and the class label c, and it can directly increase the mutual information between zs and the label c. Our model can also be extended to GAN by adding a discriminator in the pixel domain so that it produces high quality and diverse images.
****

Spectral Reconstruction From Dispersive Blur: A Novel Light Efficient Spectral Imager
Yuanyuan Zhao, Xuemei Hu, Hui Guo, Zhan Ma, Tao Yue, Xun Cao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12202-12211
Developing high light efficiency imaging techniques to retrieve high dimensional optical signal is a long-term goal in computational photography. Multispectral imaging, which captures images of different wavelengths and boosting the abilities for revealing scene properties, has developed rapidly in the last few decades. From scanning method to snapshot imaging, the limit of light collection efficiency is kept being pushed which enables wider applications especially under the light-starved scenes. In this work, we propose a novel multispectral imaging technique, that could capture the multispectral images with a high light efficiency. Through investigating the dispersive blur caused by spectral dispersers and introducing the difference of blur (DoB) constraints, we propose a basic theory for capturing multispectral information from a single dispersive-blurred image and an additional spectrum of an arbitrary point in the scene. Based on the theory, we design a prototype system and develop an optimization algorithm to realize snapshot multispectral imaging. The effectiveness of the proposed method is verified on both the synthetic data and real captured images.
****

Quasi-Unsupervised Color Constancy
Simone Bianco, Claudio Cusano; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12212-12221
We present here a method for computational color constancy in which a deep convolutional neural network is trained to detect achromatic pixels in color images after they have been converted to grayscale. The method does not require any information about the illuminant in the scene and relies on the weak assumption, fulfilled by almost all images available on the web, that training images have been

approximately balanced. Because of this requirement we define our method as qua si-unsupervised. After training, unbalanced images can be processed thanks to th e preliminary conversion to grayscale of the input to the neural network. The r esults of an extensive experimentation demonstrate that the proposed method is a ble to outperform the other unsupervised methods in the state of the art being, at the same time, flexible enough to be supervisedly fine-tuned to reach perform ance comparable with those of the best supervised methods.
****

Deep Defocus Map Estimation Using Domain Adaptation
Junyong Lee, Sungkil Lee, Sunghyun Cho, Seungyong Lee; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12 222-12230
In this paper, we propose the first end-to-end convolutional neural network (CNN) architecture, Defocus Map Estimation Network (DMENet), for spatially varying d efocus map estimation. To train the network, we produce a novel depth-of-field (DOF) dataset, SYNDOF, where each image is synthetically blurred with a ground-tr uth depth map. Due to the synthetic nature of SYNDOF, the feature characteristic s of images in SYNDOF can differ from those of real defocused photos. To address this gap, we use domain adaptation that transfers the features of real defocuse d photos into those of synthetically blurred ones. Our DMENet consists of four s ubnetworks: blur estimation, domain adaptation, content preservation, and sharpn ess calibration networks. The subnetworks are connected to each other and jointl y trained with their corresponding supervisions in an end-to-end manner. Our met hod is evaluated on publicly available blur detection and blur estimation datase ts and the results show the state-of-the-art performance.In this paper, we propo se the first end-to-end convolutional neural network (CNN) architecture, Defocus Map Estimation Network (DMENet), for spatially varying defocus map estimation. To train the network, we produce a novel depth-of-field (DOF) dataset, SYNDOF, w here each image is synthetically blurred with a ground-truth depth map. Due to t he synthetic nature of SYNDOF, the feature characteristics of images in SYNDOF c an differ from those of real defocused photos. To address this gap, we use domai n adaptation that transfers the features of real defocused photos into those of synthetically blurred ones. Our DMENet consists of four subnetworks: blur estima tion, domain adaptation, content preservation, and sharpness calibration network s. The subnetworks are connected to each other and jointly trained with their co rresponding supervisions in an end-to-end manner. Our method is evaluated on pub licly available blur detection and blur estimation datasets and the results show the state-of-the-art performance.
****

Using Unknown Occluders to Recover Hidden Scenes
Adam B. Yedidia, Manel Baradad, Christos Thrampoulidis, William T. Freeman, Gregory W. Wornell; Proceedings of the IEEE/CVF Conference on Computer Vision an d Pattern Recognition (CVPR), 2019, pp. 12231-12239
We consider the challenging problem of inferring a hidden moving scene from fain t shadows cast on a diffuse surface. Recent work in passive non-line-of-sight (N LoS) imaging has shown that the presence of occluding objects in between the sce ne and the diffuse surface significantly improves the conditioning of the proble m. However, that work assumes that the shape of the occluder is known a priori. In this paper, we relax this often impractical assumption, extending the range o f applications for passive occluder-based NLoS imaging systems. We formulate th e task of jointly recovering the unknown scene and unknown occluder as a blind d econvolution problem, for which we propose a simple but effective two-step algor ithm. At the first step, the algorithm exploits motion in the scene in order to obtain an estimate of the occluder. In particular, it exploits the fact that mot ion in realistic scenes is typically sparse. The second step is more standard: u sing regularization, we deconvolve by the occluder estimate to solve for the hid den scene. We demonstrate the effectiveness of our method with simulations and experiments in a variety of settings.
****

Competitive Collaboration: Joint Unsupervised Learning of Depth, Camera Motion,

Optical Flow and Motion Segmentation
Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas W ulff, Michael J. Black; Proceedings of the IEEE/CVF Conference on Computer Visi on and Pattern Recognition (CVPR), 2019, pp. 12240-12249
We address the unsupervised learning of several interconnected problems in low-l evel vision: single view depth prediction, camera motion estimation, optical flo w, and segmentation of a video into the static scene and moving regions. Our key insight is that these four fundamental vision problems are coupled through geom etric constraints. Consequently, learning to solve them together simplifies the problem because the solutions can reinforce each other. We go beyond previous wo rk by exploiting geometry more explicitly and segmenting the scene into static a nd moving regions. To that end, we introduce Competitive Collaboration, a framew ork that facilitates the coordinated training of multiple specialized neural net works to solve complex problems. Competitive Collaboration works much like expec tation-maximization, but with neural networks that act as both competitors to ex plain pixels that correspond to static or moving regions, and as collaborators t hrough a moderator that assigns pixels to be either static or independently movi ng. Our novel method integrates all these problems in a common framework and sim ultaneously reasons about the segmentation of the scene into moving objects and the static background, the camera motion, depth of the static scene structure, a nd the optical flow of moving objects. Our model is trained without any supervis ion and achieves state-of-the-art performance among joint unsupervised methods o n all sub-problems.
****
Learning Parallax Attention for Stereo Image Super-Resolution
Longguang Wang, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, We i An, Yulan Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12250-12259
Stereo image pairs can be used to improve the performance of super-resolution (S R) since additional information is provided from a second viewpoint. However, it is challenging to incorporate this information for SR since disparities between stereo images vary significantly. In this paper, we propose a parallax-attentio n stereo superresolution network (PASSRnet) to integrate the information from a stereo image pair for SR. Specifically, we introduce a parallax-attention mechan ism with a global receptive field along the epipolar line to handle different st ereo images with large disparity variations. We also propose a new and the large st dataset for stereo image SR (namely, Flickr1024). Extensive experiments demon strate that the parallax-attention mechanism can capture correspondence between stereo images to improve SR performance with a small computational and memory co st. Comparative results show that our PASSRnet achieves the state-of-the-art per formance on the Middlebury, KITTI 2012 and KITTI 2015 datasets.
****
Knowing When to Stop: Evaluation and Verification of Conformity to Output-Size S pecifications
Chenglong Wang, Rudy Bunel, Krishnamurthy Dvijotham, Po-Sen Huang, Edward Gr efenstette, Pushmeet Kohli; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12260-12269
Neural architectures able to generate variable-length outputs are extremely effe ctive for applications like Machine Translation and Image Captioning. In this pa per, we study the vulnerability of these models to attacks aimed at changing the output-size that can have undesirable consequences including increased computat ion and inducing faults in downstream modules that expect outputs of a certain l ength. We show the existence and construction of such attacks with two key contr ibutions. First, to overcome the difficulties of discrete search space and the n on-differentiable adversarial objective function, we develop an easy-to-compute differentiable proxy objective that can be used with gradient-based algorithms t o find output-lengthening inputs. Second, we develop a verification approach to formally prove that the network cannot produce outputs greater than a certain le ngth. Experimental results on Machine Translation and Image Captioning models s how that our adversarial output-lengthening approach can produce outputs that ar

e 50 times longer than the input, while our verification approach can, given a m odel and input domain, prove that the output length is below a certain size.
****

Spatial Attentive Single-Image Deraining With a High Quality Real Rain Dataset
Tianyu Wang, Xin Yang, Ke Xu, Shaozhe Chen, Qiang Zhang, Rynson W.H. Lau; P roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12270-12279
Removing rain streaks from a single image has been drawing considerable attentio n as rain streaks can severely degrade the image quality and affect the performa nce of existing outdoor vision tasks. While recent CNN-based derainers have repo rted promising performances, deraining remains an open problem for two reasons. First, existing synthesized rain datasets have only limited realism, in terms of modeling real rain characteristics such as rain shape, direction and intensity. Second, there are no public benchmarks for quantitative comparisons on real rai n images, which makes the current evaluation less objective. The core challenge is that real world rain/clean image pairs cannot be captured at the same time. I n this paper, we address the single image rain removal problem in two ways. Firs t, we propose a semi-automatic method that incorporates temporal priors and huma n supervision to generate a high-quality clean image from each input sequence of real rain images. Using this method, we construct a large-scale dataset of 29. 5K rain/rain-free image pairs that covers a wide range of natural rain scenes. S econd, to better cover the stochastic distribution of real rain streaks, we prop ose a novel SPatial Attentive Network (SPANet) to remove rain streaks in a local -to-global manner. Extensive experiments demonstrate that our network performs f avorably against the state-of-the-art deraining methods.
****

Focus Is All You Need: Loss Functions for Event-Based Vision
Guillermo Gallego, Mathias Gehrig, Davide Scaramuzza; Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1228 0-12289
Event cameras are novel vision sensors that output pixel-level brightness change s ("events") instead of traditional video frames. These asynchronous sensors off er several advantages over traditional cameras, such as, high temporal resolutio n, very high dynamic range, and no motion blur. To unlock the potential of such sensors, motion compensation methods have been recently proposed. We present a c ollection and taxonomy of twenty two objective functions to analyze event alignm ent in motion compensation approaches. We call them focus loss functions since t hey have strong connections with functions used in traditional shape-from-focus applications. The proposed loss functions allow bringing mature computer vision tools to the realm of event cameras. We compare the accuracy and runtime perform ance of all loss functions on a publicly available dataset, and conclude that th e variance, the gradient and the Laplacian magnitudes are among the best loss fu nctions. The applicability of the loss functions is shown on multiple tasks: rot ational motion, depth and optical flow estimation. The proposed focus loss funct ions allow to unlock the outstanding properties of event cameras.
****

Scalable Convolutional Neural Network for Image Compressed Sensing
Wuzhen Shi, Feng Jiang, Shaohui Liu, Debin Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12290-12 299
Recently, deep learning based image Compressed Sensing (CS) methods have been pr oposed and demonstrated superior reconstruction quality with low computational c omplexity. However, the existing deep learning based image CS methods need to tr ain different models for different sampling ratios, which increases the complexi ty of the encoder and decoder. In this paper, we propose a scalable convolutiona l neural network (dubbed SCSNet) to achieve scalable sampling and scalable recon struction with only one model. Specifically, SCSNet provides both coarse and fin e granular scalability. For coarse granular scalability, SCSNet is designed as a single sampling matrix plus a hierarchical reconstruction network that contains a base layer plus multiple enhancement layers. The base layer provides the basi

c reconstruction quality, while the enhancement layers reference the lower recon
struction layers and gradually improve the reconstruction quality. For fine gran
ular scalability, SCSNet achieves sampling and reconstruction at any sampling ra
tio by using a greedy method to select the measurement bases. Compared with the
existing deep learning based image CS methods, SCSNet achieves scalable sampling
 and quality scalable reconstruction at any sampling ratio with only one model.
Experimental results demonstrate that SCSNet has the state-of-the-art performanc
e while maintaining a comparable running speed with the existing deep learning b
ased image CS methods.
****

Event Cameras, Contrast Maximization and Reward Functions: An Analysis
Timo Stoffregen, Lindsay Kleeman; Proceedings of the IEEE/CVF Conference on Com
puter Vision and Pattern Recognition (CVPR), 2019, pp. 12300-12308
Event cameras asynchronously report timestamped changes in pixel intensity and o
ffer advantages over conventional raster scan cameras in terms of low-latency, l
ow redundancy sensing and high dynamic range. In recent years, much of research
in event based vision has been focused on performing tasks such as optic flow es
timation, moving object segmentation, feature tracking, camera rotation estimati
on and more, through contrast maximization. In contrast maximization, events are
 warped along motion trajectories whose parameters depend on the quantity being
estimated, to some time t_ref. The parameters are then scored by some reward fun
ction of the accumulated events at t_ref. The versatility of this approach has l
ead to a flurry of research in recent years, but no in-depth study of the reward
 chosen during optimization has yet been made. In this work we examine the choic
e of reward used in contrast maximization, propose a classification of different
 rewards and show how a reward can be constructed that is more robust to noise a
nd aperture uncertainty. We validate our work experimentally by predicting optic
al flow and comparing to ground-truth data.
****

Convolutional Neural Networks Can Be Deceived by Visual Illusions
Alexander Gomez-Villa, Adrian Martin, Javier Vazquez-Corral, Marcelo Bertalmi
o; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni
tion (CVPR), 2019, pp. 12309-12317
Visual illusions teach us that what we see is not always what is represented in
the physical world. Their special nature make them a fascinating tool to test an
d validate any new vision model proposed. In general, current vision models are
based on the concatenation of linear and non-linear operations. The similarity o
f this structure with the operations present in Convolutional Neural Networks (C
NNs) has motivated us to study if CNNs trained for low-level visual tasks are de
ceived by visual illusions. In particular, we show that CNNs trained for image d
enoising, image deblurring, and computational color constancy are able to replic
ate the human response to visual illusions, and that the extent of this replicat
ion varies with respect to variation in architecture and spatial pattern size.
These results suggest that in order to obtain CNNs that better replicate human b
ehaviour, we may need to start aiming for them to better replicate visual illusi
ons.
****

PDE Acceleration for Active Contours
Anthony Yezzi, Ganesh Sundaramoorthi, Minas Benyamin; Proceedings of the IEEE/
CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1231
8-12328
Following the seminal work of Nesterov, accelerated optimization methods have
been used to powerfully boost the performance of first-order, gradient-based p
arameter estimation in scenarios where second-order optimization strategies ar
e either inapplicable or impractical. Accelerated gradient descent converges
faster and performs a more robust local search of the parameter space by ini
tially overshooting then oscillating back into minimizers which have a basis o
f attraction large enough to contain the overshoot. Recent work has demonstrat
ed how a broad class of accelerated schemes can be cast in a variational frame
work leading to continuum limit ODE's. We extend their formulation to the PDE

framework, specifically for the infinite dimensional manifold of continuous curves, to introduce acceleration, and its added robustness, into the broad range of PDE based active contours.
****

Dichromatic Model Based Temporal Color Constancy for AC Light Sources
Jun-Sang Yoo, Jong-Ok Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12329-12338
Existing dichromatic color constancy approach commonly requires a number of spatial pixels which have high specularity. In this paper, we propose a novel approach to estimate the illuminant chromaticity of AC light source using high-speed camera. We found that the temporal observations of an image pixel at a fixed location distribute on an identical dichromatic plane. Instead of spatial pixels with high specularity, multiple temporal samples of a pixel are exploited to determine AC pixels for dichromatic plane estimation, whose pixel intensity is sinusoidally varying well. A dichromatic plane is calculated per each AC pixel, and illuminant chromaticity is determined by the intersection of dichromatic planes. From multiple dichromatic planes, an optimal illuminant is estimated with a novel MAP framework. It is shown that the proposed method outperforms both existing dichromatic based methods and temporal color constancy methods, irrespective of the amount of specularity.
****

Semantic Attribute Matching Networks
Seungryong Kim, Dongbo Min, Somi Jeong, Sunok Kim, Sangryul Jeon, Kwanghoon Sohn; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12339-12348
We present semantic attribute matching networks (SAM-Net) for jointly establishing correspondences and transferring attributes across semantically similar images, which intelligently weaves the advantages of the two tasks while overcoming their limitations. SAM-Net accomplishes this through an iterative process of establishing reliable correspondences by reducing the attribute discrepancy between the images and synthesizing attribute transferred images using the learned correspondences. To learn the networks using weak supervisions in the form of image pairs, we present a semantic attribute matching loss based on the matching similarity between an attribute transferred source feature and a warped target feature. With SAM-Net, the state-of-the-art performance is attained on several benchmarks for semantic matching and attribute transfer.
****

Skin-Based Identification From Multispectral Image Data Using CNNs
Takeshi Uemori, Atsushi Ito, Yusuke Moriuchi, Alexander Gatto, Jun Murayama; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12349-12358
User identification from hand images only is still a challenging task. In this paper, we propose a new biometric identification system based solely on a skin patch from a multispectral image. The system is utilizing a novel modified 3D CNN architecture which is taking advantage of multispectral data. We demonstrate the application of our system for the example of human identification from multispectral images of hands. To the best of our knowledge, this paper is the first to describe a pose-invariant and robust to overlapping real-time human identification system using hands. Additionally, we provide a framework to optimize the required spectral bands for the given spatial resolution limitations.
****

Large-Scale Distributed Second-Order Optimization Using Kronecker-Factored Approximate Curvature for Deep Convolutional Neural Networks
Kazuki Osawa, Yohei Tsuji, Yuichiro Ueno, Akira Naruse, Rio Yokota, Satoshi Matsuoka; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12359-12367
Large-scale distributed training of deep neural networks suffers from the generalization gap caused by the increase in the effective mini-batch size. Previous approaches try to solve this problem by varying the learning rate and batch size over epochs and layers, or some ad hoc modification of the batch normalization.

We propose an alternative approach using a second order optimization method that shows similar generalization capability to first order methods, but converges faster and can handle larger mini-batches. To test our method on a benchmark where highly optimized first order methods are available as references, we train ResNet-50 on ImageNet-1K. We converged to 75% Top-1 validation accuracy in 35 epochs for mini-batch sizes under 16,384, and achieved 75% even with a mini-batch size of 131,072, which took only 978 iterations.
****

Putting Humans in a Scene: Learning Affordance in 3D Indoor Environments
Xueting Li, Sifei Liu, Kihwan Kim, Xiaolong Wang, Ming-Hsuan Yang, Jan Kautz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12368-12376
Affordance modeling plays an important role in visual understanding. In this paper, we aim to predict affordances of 3D indoor scenes, specifically what human poses are afforded by a given indoor environment, such as sitting on a chair or standing on the floor. In order to predict valid affordances and learn possible 3D human poses in indoor scenes, we need to understand the semantic and geometric structure of a scene as well as its potential interactions with a human. To learn such a model, a large-scale dataset of 3D indoor affordances is required. In this work, we build a fully automatic 3D pose synthesizer that fuses semantic knowledge from a large number of 2D poses extracted from TV shows as well as 3D geometric knowledge from voxel representations of indoor scenes. With the data created by the synthesizer, we introduce a 3D pose generative model to predict semantically plausible and physically feasible human poses within a given scene (provided as a single RGB, RGB-D, or depth image). We demonstrate that our human affordance prediction method consistently outperforms existing state-of-the-art methods.
****

PIEs: Pose Invariant Embeddings
Chih-Hui Ho, Pedro Morgado, Amir Persekian, Nuno Vasconcelos; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12377-12386
The role of pose invariance in image recognition and retrieval is studied. A taxonomic classification of embeddings, according to their level of invariance, is introduced and used to clarify connections between existing embeddings, identify missing approaches, and propose invariant generalizations. This leads to a new family of pose invariant embeddings (PIEs), derived from existing approaches by a combination of two models, which follow from the interpretation of CNNs as estimators of class posterior probabilities: a view-to-object model and an object-to-class model. The new pose-invariant models are shown to have interesting properties, both theoretically and through experiments, where they outperform existing multiview approaches. Most notably, they achieve good performance for both 1) classification and retrieval, and 2) single and multiview inference. These are important properties for the design of real vision systems, where universal embeddings are preferable to task specific ones, and multiple images are usually not available at inference time. Finally, a new multiview dataset of real objects, imaged in the wild against complex backgrounds, is introduced. We believe that this is a much needed complement to the synthetic datasets in wide use and will contribute to the advancement of multiview recognition and retrieval.
****

Representation Similarity Analysis for Efficient Task Taxonomy & Transfer Learning
Kshitij Dwivedi, Gemma Roig; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12387-12396
Transfer learning is widely used in deep neural network models when there are few labeled examples available. The common approach is to take a pre-trained network in a similar task and finetune the model parameters. This is usually done blindly without a pre-selection from a set of pre-trained models, or by finetuning a set of models trained on different tasks and selecting the best performing one by cross-validation. We address this problem by proposing an approach to asses

s the relationship between visual tasks and their task-specific models. Our meth od uses Representation Similarity Analysis (RSA), which is commonly used to find a correlation between neuronal responses from brain data and models. With RSA w e obtain a similarity score among tasks by computing correlations between models trained on different tasks. Our method is efficient as it requires only pre-tra ined models, and a few images with no further training. We demonstrate the effec tiveness and efficiency of our method to generating task taxonomy on Taskonomy d ataset. We next evaluate the relationship of RSA with the transfer learning per formance on Taskonomy tasks and a new task: Pascal VOC semantic segmentation. Ou r results reveal that models trained on tasks with higher similarity score show higher transfer learning performance. Surprisingly, the best transfer learning r esult for Pascal VOC semantic segmentation is not obtained from the pre-trained model on semantic segmentation, probably due to the domain differences, and our method successfully selects the high performing models.

****

Object Counting and Instance Segmentation With Image-Level Supervision

Hisham Cholakkal, Guolei Sun, Fahad Shahbaz Khan, Ling Shao; Proceedings of t he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12397-12405

Common object counting in a natural scene is a challenging problem in computer v ision with numerous real-world applications. Existing image-level supervised com mon object counting approaches only predict the global object count and rely on additional instance-level supervision to also determine object locations. We pro pose an image-level supervised approach that provides both the global object cou nt and the spatial distribution of object instances by constructing an object ca tegory density map. Motivated by psychological studies, we further reduce image- level supervision using a limited object count information (up to four). To the best of our knowledge, we are the first to propose image-level supervised densit y map estimation for common object counting and demonstrate its effectiveness in image-level supervised instance segmentation. Comprehensive experiments are pe rformed on the PASCAL VOC and COCO datasets. Our approach outperforms existing m ethods, including those using instance-level supervision, on both datasets for c ommon object counting. Moreover, our approach improves state-of-the-art image-l evel supervised instance segmentation with a relative gain of 17.8% in terms of average best overlap, on the PASCAL VOC 2012 dataset.

****

Variational Autoencoders Pursue PCA Directions (by Accident)

Michal Rolinek, Dominik Zietlow, Georg Martius; Proceedings of the IEEE/CVF Co nference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12406-1241 5

The Variational Autoencoder (VAE) is a powerful architecture capable of represen tation learning and generative modeling. When it comes to learning interpretable (disentangled) representations, VAE and its variants show unparalleled performa nce. However, the reasons for this are unclear, since a very particular alignmen t of the latent embedding is needed but the design of the VAE does not encourage it in any explicit way. We address this matter and offer the following explanat ion: the diagonal approximation in the encoder together with the inherent stocha sticity force local orthogonality of the decoder. The local behavior of promotin g both reconstruction and orthogonality matches closely how the PCA embedding is chosen. Alongside providing an intuitive understanding, we justify the statemen t with full theoretical analysis as well as with experiments.

****

A Relation-Augmented Fully Convolutional Network for Semantic Segmentation in Ae rial Scenes

Lichao Mou, Yuansheng Hua, Xiao Xiang Zhu; Proceedings of the IEEE/CVF Confere nce on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12416-12425

Most current semantic segmentation approaches fall back on deep convolutional ne ural networks (CNNs). However, their use of convolution operations with local re ceptive fields causes failures in modeling contextual spatial relations. Prior w orks have sought to address this issue by using graphical models or spatial prop

agation modules in networks. But such models often fail to capture long-range spatial relationships between entities, which leads to spatially fragmented predictions. Moreover, recent works have demonstrated that channel-wise information also acts a pivotal part in CNNs. In this work, we introduce two simple yet effective network units, the spatial relation module and the channel relation module, to learn and reason about global relationships between any two spatial positions or feature maps, and then produce relation-augmented feature representations. The spatial and channel relation modules are general and extensible, and can be used in a plug-and-play fashion with the existing fully convolutional network (FCN) framework. We evaluate relation module-equipped networks on semantic segmentation tasks using two aerial image datasets, which fundamentally depend on long-range spatial relational reasoning. The networks achieve very competitive results, bringing significant improvements over baselines.
****

Temporal Transformer Networks: Joint Learning of Invariant and Discriminative Time Warping
Suhas Lohit, Qiao Wang, Pavan Turaga; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12426-12435
Many time-series classification problems involve developing metrics that are invariant to temporal misalignment. In human activity analysis, temporal misalignment arises due to various reasons including differing initial phase, sensor sampling rates, and elastic time-warps due to subject-specific biomechanics. Past work in this area has only looked at reducing intra-class variability by elastic temporal alignment. In this paper, we propose a hybrid model-based and data-driven approach to learn warping functions that not just reduce intra-class variability, but also increase inter-class separation. We call this a temporal transformer network (TTN). TTN is an interpretable differentiable module, which can be easily integrated at the front end of a classification network. The module is capable of reducing intra-class variance by generating input-dependent warping functions which lead to rate-robust representations. At the same time, it increases inter-class variance by learning warping functions that are more discriminative. We show improvements over strong baselines in 3D action recognition on challenging datasets using the proposed framework. The improvements are especially pronounced when training sets are smaller.
****

PCAN: 3D Attention Map Learning Using Contextual Information for Point Cloud Based Retrieval
Wenxiao Zhang, Chunxia Xiao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12436-12445
Point cloud based retrieval for place recognition is an emerging problem in vision field. The main challenge is how to find an efficient way to encode the local features into a discriminative global descriptor. In this paper, we propose a Point Contextual Attention Network (PCAN), which can predict the significance of each local point feature based on point context. Our network makes it possible to pay more attention to the task-relevent features when aggregating local features. Experiments on various benchmark datasets show that the proposed network can provide outperformance than current state-of-the-art approaches.
****

Depth Coefficients for Depth Completion
Saif Imran, Yunfei Long, Xiaoming Liu, Daniel Morris; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12446-12455
Depth completion involves estimating a dense depth image from sparse depth measurements, often guided by a color image. While linear upsampling is straight forward, it results in depth pixels being interpolated in empty space across discontinuities between objects. Current methods use deep networks to maintain gaps between objects. Nevertheless depth smearing remains a challenge. We propose a new representation for depth called Depth Coefficients (DC) to address this problem. It enables convolutions to more easily avoid inter-object depth mixing. We also show that the standard Mean Squared Error (MSE) loss function can promote

depth mixing, and so we propose instead to use cross-entropy loss for DC. Both q
uantitative and qualitative evaluation are conducted on benchmarks, and we show
that switching out sparse depth input and MSE loss functions with our DC represe
ntation and loss is a simple way to improve performance, reduce pixel depth mixi
ng and can improve object detection.
****

Diversify and Match: A Domain Adaptive Representation Learning Paradigm for Obje
ct Detection
Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, Changick Kim; Proce
edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV
PR), 2019, pp. 12456-12465
We introduce a novel unsupervised domain adaptation approach for object detectio
n. We aim to alleviate the imperfect translation problem of pixel-level adaptati
ons, and the source-biased discriminativity problem of feature-level adaptations
simultaneously. Our approach is composed of two stages, i.e., Domain Diversific
ation (DD) and Multi-domain-invariant Representation Learning (MRL). At the DD s
tage, we diversify the distribution of the labeled data by generating various di
stinctive shifted domains from the source domain. At the MRL stage, we apply adv
ersarial learning with a multi-domain discriminator to encourage feature to be i
ndistinguishable among the domains. DD addresses the source-biased discriminativ
ity, while MRL mitigates the imperfect image translation. We construct a structu
red domain adaptation framework for our learning paradigm and introduce a practi
cal way of DD for implementation. Our method outperforms the state-of-the-art me
thods by a large margin of 3% 11% in terms of mean average precision (mAP) on va
rious datasets.
****

Good News, Everyone! Context Driven Entity-Aware Captioning for News Images
Ali Furkan Biten, Lluis Gomez, Marcal Rusinol, Dimosthenis Karatzas; Proceedi
ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
, 2019, pp. 12466-12475
Current image captioning systems perform at a merely descriptive level, essentia
lly enumerating the objects in the scene and their relations. Humans, on the con
trary, interpret images by integrating several sources of prior knowledge of the
world. In this work, we aim to take a step closer to producing captions that of
fer a plausible interpretation of the scene, by integrating such contextual info
rmation into the captioning pipeline. For this we focus on the captioning of ima
ges used to illustrate news articles. We propose a novel captioning method that
is able to leverage contextual information provided by the text of news articles
associated with an image. Our model is able to selectively draw information fro
m the article guided by visual cues, and to dynamically extend the output dictio
nary to out-of-vocabulary named entities that appear in the context source. Furt
hermore we introduce "GoodNews", the largest news image captioning dataset in th
e literature and demonstrate state-of-the-art results.
****

Multi-Level Multimodal Common Semantic Space for Image-Phrase Grounding
Hassan Akbari, Svebor Karaman, Surabhi Bhargava, Brian Chen, Carl Vondrick,
Shih-Fu Chang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pa
ttern Recognition (CVPR), 2019, pp. 12476-12486
We address the problem of phrase grounding by learning a multi-level common sema
ntic space shared by the textual and visual modalities. We exploit multiple leve
ls of feature maps of a Deep Convolutional Neural Network, as well as contextual
ized word and sentence embeddings extracted from a character-based language mode
l. Following dedicated non-linear mappings for visual features at each level, wo
rd, and sentence embeddings, we obtain multiple instantiations of our common sem
antic space in which comparisons between any target text and the visual content
is performed with cosine similarity. We guide the model by a multi-level multimo
dal attention mechanism which outputs attended visual features at each level. Th
e best level is chosen to be compared with text content for maximizing the perti
nence scores of image-sentence pairs of the ground truth. Experiments conducted
on three publicly available datasets show significant performance gains (20%-60%

relative) over the state-of-the-art in phrase localization and set a new performance record on those datasets. We provide a detailed ablation study to show the contribution of each element of our approach and release our code on GitHub.
****

Spatio-Temporal Dynamics and Semantic Attribute Enriched Visual Encoding for Video Captioning
Nayyer Aafaq, Naveed Akhtar, Wei Liu, Syed Zulqarnain Gilani, Ajmal Mian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12487-12496
Automatic generation of video captions is a fundamental challenge in computer vision. Recent techniques typically employ a combination of Convolutional Neural Networks (CNNs) and Recursive Neural Networks (RNNs) for video captioning. These methods mainly focus on tailoring sequence learning through RNNs for better caption generation, whereas off-the-shelf visual features are borrowed from CNNs. We argue that careful designing of visual features for this task is equally important, and present a visual feature encoding technique to generate semantically rich captions using Gated Recurrent Units (GRUs). Our method embeds rich temporal dynamics in visual features by hierarchically applying Short Fourier Transform to CNN features of the whole video. It additionally derives high level semantics from an object detector to enrich the representation with spatial dynamics of the detected objects. The final representation is projected to a compact space and fed to a language model. By learning a relatively simple language model comprising two GRU layers, we establish new state-of-the-art on MSVD and MSR-VTT datasets for METEOR and ROUGE_L metrics.
****

Pointing Novel Objects in Image Captioning
Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, Tao Mei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12497-12506
Image captioning has received significant attention with remarkable improvements in recent advances. Nevertheless, images in the wild encapsulate rich knowledge and cannot be sufficiently described with models built on image-caption pairs containing only in-domain objects. In this paper, we propose to address the problem by augmenting standard deep captioning architectures with object learners. Specifically, we present Long Short-Term Memory with Pointing (LSTM-P) --- a new architecture that facilitates vocabulary expansion and produces novel objects via pointing mechanism. Technically, object learners are initially pre-trained on available object recognition data. Pointing in LSTM-P then balances the probability between generating a word through LSTM and copying a word from the recognized objects at each time step in decoder stage. Furthermore, our captioning encourages global coverage of objects in the sentence. Extensive experiments are conducted on both held-out COCO image captioning and ImageNet datasets for describing novel objects, and superior results are reported when comparing to state-of-the-art approaches. More remarkably, we obtain an average of 60.9% in F1 score on held-out COCO dataset.
****

Informative Object Annotations: Tell Me Something I Don't Know
Lior Bracha, Gal Chechik; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12507-12515
Capturing the interesting components of an image is a key aspect of image understanding. When a speaker annotates an image, selecting labels that are informative greatly depends on the prior knowledge of a prospective listener. Motivated by cognitive theories of categorization and communication, we present a new unsupervised approach to model this prior knowledge and quantify the informativeness of a description. Specifically, we compute how knowledge of a label reduces uncertainty over the space of labels and use this uncertainty reduction to rank candidate labels for describing an image. While the full estimation problem is intractable, we describe an efficient algorithm to approximate entropy reduction using a tree-structured graphical model. We evaluate our approach on the open-images dataset using a new evaluation set of 10K ground-truth ratings and find that it

achieves over 65% agreement with human raters, close to the upper bound of inter-rater agreement and largely outperforming other unsupervised baseline approaches.
****

Engaging Image Captioning via Personality
Kurt Shuster, Samuel Humeau, Hexiang Hu, Antoine Bordes, Jason Weston; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12516-12526
Standard image captioning tasks such as COCO and Flickr30k are factual, neutral in tone and (to a human) state the obvious (e.g., "a man playing a guitar"). While such tasks are useful to verify that a machine understands the content of an image, they are not engaging to humans as captions. With this in mind we define a new task, PERSONALITY-CAPTIONS, where the goal is to be as engaging to humans as possible by incorporating controllable style and personality traits. We collect and release a large dataset of 241,858 of such captions conditioned over 215 possible traits. We build models that combine existing work from (i) sentence representations [36] with Transformers trained on 1.7 billion dialogue examples; and (ii) image representations [32] with ResNets trained on 3.5 billion social media images. We obtain state-of-the-art performance on Flickr30k and COCO, and strong performance on our new task. Finally, online evaluations validate that our task and models are engaging to humans, with our best model close to human performance.
****

Vision-Based Navigation With Language-Based Assistance via Imitation Learning With Indirect Intervention
Khanh Nguyen, Debadeepta Dey, Chris Brockett, Bill Dolan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12527-12537
We present Vision-based Navigation with Language-based Assistance (VNLA), a grounded vision-language task where an agent with visual perception is guided via language to find objects in photorealistic indoor environments. The task emulates a real-world scenario in that (a) the requester may not know how to navigate to the target objects and thus makes requests by only specifying high-level end-goals, and (b) the agent is capable of sensing when it is lost and querying an advisor, who is more qualified at the task, to obtain language subgoals to make progress. To model language-based assistance, we develop a general framework termed Imitation Learning with Indirect Intervention (I3L), and propose a solution that is effective on the VNLA task. Empirical results show that this approach significantly improves the success rate of the learning agent over other baselines on both seen and unseen environments. Our code and data are publicly available at https://github.com/debadeepta/vnla .
****

TOUCHDOWN: Natural Language Navigation and Spatial Reasoning in Visual Street Environments
Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, Yoav Artzi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12538-12547
We study the problem of jointly reasoning about language and vision through a navigation and spatial reasoning task. We introduce the Touchdown task and dataset, where an agent must first follow navigation instructions in a Street View environment to a goal position, and then guess a location in its observed environment described in natural language to find a hidden object. The data contains 9326 examples of English instructions and spatial descriptions paired with demonstrations. We perform qualitative linguistic analysis, and show that the data displays a rich use of spatial reasoning. Empirical analysis shows the data presents an open challenge to existing methods.
****

A Simple Baseline for Audio-Visual Scene-Aware Dialog
Idan Schwartz, Alexander G. Schwing, Tamir Hazan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12548-12

558
The recently proposed audio-visual scene-aware dialog task paves the way to a more data-driven way of learning virtual assistants, smart speakers and car navigation systems. However, very little is known to date about how to effectively extract meaningful information from a plethora of sensors that pound the computational engine of those devices. Therefore, in this paper, we provide and carefully analyze a simple baseline for audio-visual scene-aware dialog which is trained end-to-end. Our method differentiates in a data-driven manner useful signals from distracting ones using an attention mechanism. We evaluate the proposed approach on the recently introduced and challenging audio-visual scene-aware dataset, and demonstrate the key features that permit to outperform the current state-of-the-art by more than 20% on CIDEr.
****
End-To-End Learned Random Walker for Seeded Image Segmentation
Lorenzo Cerrone, Alexander Zeilmann, Fred A. Hamprecht; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12559-12568
We present an end-to-end learned algorithm for seeded segmentation. Our method is based on the Random Walker algorithm, where we predict the edge weights of the un- derlying graph using a convolutional neural network. This can be interpreted as learning context-dependent diffusiv- ities for a linear diffusion process. After calculating the exact gradient for optimizing these diffusivities, we pro- pose simplifications that sparsely sample the gradient while still maintaining competitive results. The proposed method achieves the currently best results on the seeded CREMI neuron segmentation challenge.
****
Efficient Neural Network Compression
Hyeji Kim, Muhammad Umar Karim Khan, Chong-Min Kyung; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12569-12577
Network compression reduces the computational complexity and memory consumption of deep neural networks by reducing the number of parameters. In SVD-based network compression the right rank needs to be decided for every layer of the network. In this paper we propose an efficient method for obtaining the rank configuration of the whole network. Unlike previous methods which consider each layer separately, our method considers the whole network to choose the right rank configuration. We propose novel accuracy metrics to represent the accuracy and complexity relationship for a given neural network. We use these metrics in a non-iterative fashion to obtain the right rank configuration which satisfies the constraints on FLOPs and memory while maintaining sufficient accuracy. Experiments show that our method provides better compromise between accuracy and computational complexity/memory consumption while performing compression at much higher speed. For VGG-16 our network can reduce the FLOPs by 25% and improve accuracy by 0.7% compared to the baseline, while requiring only 3 minutes on a CPU to search for the right rank configuration. Previously, similar results were achieved in 4 hours with 8 GPUs. The proposed method can be used for lossless compression of a neural network as well. The better accuracy and complexity compromise, as well as the extremely fast speed of our method make it suitable for neural network compression.
****
Cascaded Generative and Discriminative Learning for Microcalcification Detection in Breast Mammograms
Fandong Zhang, Ling Luo, Xinwei Sun, Zhen Zhou, Xiuli Li, Yizhou Yu, Yizhou Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12578-12586
Accurate microcalcification (mC) detection is of great importance due to its high proportion in early breast cancers. Most of the previous mC detection methods belong to discriminative models, where classifiers are exploited to distinguish mCs from other backgrounds. However, it is still challenging for these methods to tell the mCs from amounts of normal tissues because they are too tiny (at most

14 pixels). Generative methods can precisely model the normal tissues and regard the abnormal ones as outliers, while they fail to further distinguish the mCs from other anomalies, i.e. vessel calcifications. In this paper, we propose a hybrid approach by taking advantages of both generative and discriminative models. Firstly, a generative model named Anomaly Separation Network (ASN) is used to generate candidate mCs. ASN contains two major components. A deep convolutional encoder-decoder network is built to learn the image reconstruction mapping and a t-test loss function is designed to separate the distributions of the reconstruction residuals of mCs from normal tissues. Secondly, a discriminative model is cascaded to tell the mCs from the false positives. Finally, to verify the effectiveness of our method, we conduct experiments on both public and in-house datasets, which demonstrates that our approach outperforms previous state-of-the-art methods.

****

C3AE: Exploring the Limits of Compact Model for Age Estimation

Chao Zhang, Shuaicheng Liu, Xun Xu, Ce Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12587-12596

Age estimation is a classic learning problem in computer vision. Many larger and deeper CNNs have been proposed with promising performance, such as AlexNet, VggNet, GoogLeNet and ResNet. However, these models are not practical for the embedded/mobile devices. Recently, MobileNets and ShuffleNets have been proposed to reduce the number of parameters, yielding lightweight models. However, their representation has been weakened because of the adoption of depth-wise separable convolution. In this work, we investigate the limits of compact model for small-scale image and propose an extremely Compact yet efficient Cascade Context-based Age Estimation model(C3AE). This model possesses only 1/9 and 1/2000 parameters compared with MobileNets/ShuffleNets and VggNet, while achieves competitive performance. In particular, we re-define age estimation problem by two-points representation, which is implemented by a cascade model. Moreover, to fully utilize the facial context information, multi-branch CNN network is proposed to aggregate multi-scale context. Experiments are carried out on three age estimation datasets. The state-of-the-art performance on compact model has been achieved with a relatively large margin.

****

Adaptive Weighting Multi-Field-Of-View CNN for Semantic Segmentation in Pathology

Hiroki Tokunaga, Yuki Teramoto, Akihiko Yoshizawa, Ryoma Bise; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12597-12606

Automated digital histopathology image segmentation is an important task to help pathologists diagnose tumors and cancer subtypes. For pathological diagnosis of cancer subtypes, pathologists usually change the magnification of whole-slide images (WSI) viewers. A key assumption is that the importance of the magnifications depends on the characteristics of the input image, such as cancer subtypes. In this paper, we propose a novel semantic segmentation method, called Adaptive-Weighting-Multi-Field-of-View-CNN (AWMF-CNN), that can adaptively use image features from images with different magnifications to segment multiple cancer subtype regions in the input image. The proposed method aggregates several expert CNNs for images of different magnifications by adaptively changing the weight of each expert depending on the input image. It leverages information in the images with different magnifications that might be useful for identifying the subtypes. It outperformed other state-of-the-art methods in experiments.

****

In Defense of Pre-Trained ImageNet Architectures for Real-Time Semantic Segmentation of Road-Driving Images

Marin Orsic, Ivan Kreso, Petra Bevandic, Sinisa Segvic; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12607-12616

Recent success of semantic segmentation approaches on demanding road driving datasets has spurred interest in many related application fields. Many of these app

lications involve real-time prediction on mobile platforms such as cars, drones and various kinds of robots. Real-time setup is challenging due to extraordinary computational complexity involved. Many previous works address the challenge with custom lightweight architectures which decrease computational complexity by reducing depth, width and layer capacity with respect to general purpose architectures. We propose an alternative approach which achieves a significantly better performance across a wide range of computing budgets. First, we rely on a light-weight general purpose architecture as the main recognition engine. Then, we leverage light-weight upsampling with lateral connections as the most cost-effective solution to restore the prediction resolution. Finally, we propose to enlarge the receptive field by fusing shared features at multiple resolutions in a novel fashion. Experiments on several road driving datasets show a substantial advantage of the proposed approach, either with ImageNet pre-trained parameters or when we learn from scratch. Our Cityscapes test submission entitled SwiftNetRN-18 delivers 75.5% MIoU and achieves 39.9 Hz on 1024x2048 images on GTX1080Ti.

****

Context-Aware Visual Compatibility Prediction

Guillem Cucurull, Perouz Taslakian, David Vazquez; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12617-12626

How do we determine whether two or more clothing items are compatible or visually appealing? Part of the answer lies in understanding of visual aesthetics, and is biased by personal preferences shaped by social attitudes, time, and place. In this work we propose a method that predicts compatibility between two items based on their visual features, as well as their context. We define context as the products that are known to be compatible with each of these item. Our model is in contrast to other metric learning approaches that rely on pairwise comparisons between item features alone. We address the compatibility prediction problem using a graph neural network that learns to generate product embeddings conditioned on their context. We present results for two prediction tasks (fill in the blank and outfit compatibility) tested on two fashion datasets Polyvore and Fashion-Gen, and on a subset of the Amazon dataset; we achieve state of the art results when using context information and show how test performance improves as more context is used.

****

Sim-To-Real via Sim-To-Sim: Data-Efficient Robotic Grasping via Randomized-To-Canonical Adaptation Networks

Stephen James, Paul Wohlhart, Mrinal Kalakrishnan, Dmitry Kalashnikov, Alex Irpan, Julian Ibarz, Sergey Levine, Raia Hadsell, Konstantinos Bousmalis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12627-12637

Real world data, especially in the domain of robotics, is notoriously costly to collect. One way to circumvent this can be to leverage the power of simulation to produce large amounts of labelled data. However, training models on simulated images does not readily transfer to real-world ones. Using domain adaptation methods to cross this "reality gap" requires a large amount of unlabelled real-world data, whilst domain randomization alone can waste modeling power. In this paper, we present Randomized-to-Canonical Adaptation Networks (RCANs), a novel approach to crossing the visual reality gap that uses no real-world data. Our method learns to translate randomized rendered images into their equivalent non-randomized, canonical versions. This in turn allows for real images to also be translated into canonical sim images. We demonstrate the effectiveness of this sim-to-real approach by training a vision-based closed-loop grasping reinforcement learning agent in simulation, and then transferring it to the real world to attain 70% zero-shot grasp success on unseen objects, a result that almost doubles the success of learning the same task directly on domain randomization alone. Additionally, by joint finetuning in the real-world with only 5,000 real-world grasps, our method achieves 91%, attaining comparable performance to a state-of-the-art system trained with 580,000 real-world grasps, resulting in a reduction of real-world data by more than 99%.

****
Multiview 2D/3D Rigid Registration via a Point-Of-Interest Network for Tracking and Triangulation
Haofu Liao, Wei-An Lin, Jiarui Zhang, Jingdan Zhang, Jiebo Luo, S. Kevin Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12638-12647
We propose to tackle the problem of multiview 2D/3D rigid registration for intervention via a Point-Of-Interest Network for Tracking and Triangulation (POINT^2). POINT^2 learns to establish 2D point-to-point correspondences between the pre- and intra-intervention images by tracking a set of random POIs. The 3D pose of the pre-intervention volume is then estimated through a triangulation layer. In POINT^2, the unified framework of the POI tracker and the triangulation layer enables learning informative 2D features and estimating 3D pose jointly. In contrast to existing approaches, POINT^2 only requires a single forward-pass to achieve a reliable 2D/3D registration. As the POI tracker is shift-invariant, POINT^2 is more robust to the initial pose of the 3D pre-intervention image. Extensive experiments on a large-scale clinical cone-beam CT (CBCT) dataset show that the proposed POINT^2 method outperforms the existing learning-based method in terms of accuracy, robustness and running time. Furthermore, when used as an initial pose estimator, our method also improves the robustness and speed of the state-of-the-art optimization-based approaches by ten folds.
****
Context-Aware Spatio-Recurrent Curvilinear Structure Segmentation
Feigege Wang, Yue Gu, Wenxi Liu, Yuanlong Yu, Shengfeng He, Jia Pan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12648-12657
Curvilinear structures are frequently observed in various images in different forms, such as blood vessels or neuronal boundaries in biomedical images. In this paper, we propose a novel curvilinear structure segmentation approach using context-aware spatio-recurrent networks. Instead of directly segmenting the whole image or densely segmenting fixed-sized local patches, our method recurrently samples patches with varied scales from the target image with learned policy and processes them locally, which is similar to the behavior of changing retinal fixations in the human visual system and it is beneficial for capturing the multi-scale or hierarchical modality of the complex curvilinear structures. In specific, the policy of choosing local patches is attentively learned based on the contextual information of the image and the historical sampling experience. In this way, with more patches sampled and refined, the segmentation of the whole image can be progressively improved. To validate our approach, comparison experiments on different types of image data are conducted and the sampling procedures for exemplar images are illustrated. We demonstrate that our method achieves the state-of-the-art performance in public datasets.
****
An Alternative Deep Feature Approach to Line Level Keyword Spotting
George Retsinas, Georgios Louloudis, Nikolaos Stamatopoulos, Giorgos Sfikas, Basilis Gatos; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12658-12666
Keyword spotting (KWS) is defined as the problem of detecting all instances of a given word, provided by the user either as a query word image (Query-by-Example, QbE) or a query word string (Query-by-String, QbS) in a body of digitized documents. Keyword detection is typically preceded by a preprocessing step where the text is segmented into text lines (line-level KWS). Methods following this paradigm are monopolized by test-time computationally expensive handwritten text recognition (HTR)-based approaches; furthermore, they typically cannot handle image queries (QbE). In this work, we propose a time and storage-efficient, deep feature-based approach that enables both the image and textual search options. Three distinct components, all modeled as neural networks, are combined: normalization, feature extraction and representation of image and textual input into a common space. These components, even if designed on word level image representations, collaborate in order

to achieve an efficient line level keyword spotting system. The experimental results indicate that the proposed system is on par with state-of-the-art KWS methods.
****

Dynamics Are Important for the Recognition of Equine Pain in Video
Sofia Broome, Karina Bech Gleerup, Pia Haubro Andersen, Hedvig Kjellstrom; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12667-12676
A prerequisite to successfully alleviate pain in animals is to recognize it, which is a great challenge in non-verbal species. Furthermore, prey animals such as horses tend to hide their pain. In this study, we propose a deep recurrent two-stream architecture for the task of distinguishing pain from non-pain in videos of horses. Different models are evaluated on a unique dataset showing horses under controlled trials with moderate pain induction, which has been presented in earlier work. Sequential models are experimentally compared to single-frame models, showing the importance of the temporal dimension of the data, and are benchmarked against a veterinary expert classification of the data. We additionally perform baseline comparisons with generalized versions of state-of-the-art human pain recognition methods. While equine pain detection in machine learning is a novel field, our results surpass veterinary expert performance and outperform pain detection results reported for other larger non-human species.
****

LaserNet: An Efficient Probabilistic 3D Object Detector for Autonomous Driving
Gregory P. Meyer, Ankit Laddha, Eric Kee, Carlos Vallespi-Gonzalez, Carl K. Wellington; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12677-12686
In this paper, we present LaserNet, a computationally efficient method for 3D object detection from LiDAR data for autonomous driving. The efficiency results from processing LiDAR data in the native range view of the sensor, where the input data is naturally compact. Operating in the range view involves well known challenges for learning, including occlusion and scale variation, but it also provides contextual information based on how the sensor data was captured. Our approach uses a fully convolutional network to predict a multimodal distribution over 3D boxes for each point and then it efficiently fuses these distributions to generate a prediction for each object. Experiments show that modeling each detection as a distribution rather than a single deterministic box leads to better overall detection performance. Benchmark results show that this approach has significantly lower runtime than other recent detectors and that it achieves state-of-the-art performance when compared on a large dataset that has enough data to overcome the challenges of training on the range view.
****

Machine Vision Guided 3D Medical Image Compression for Efficient Transmission and Accurate Segmentation in the Clouds
Zihao Liu, Xiaowei Xu, Tao Liu, Qi Liu, Yanzhi Wang, Yiyu Shi, Wujie Wen, Meiping Huang, Haiyun Yuan, Jian Zhuang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12687-12696
Cloud based medical image analysis has become popular recently due to the high computation complexities of various deep neural network (DNN) based frameworks and the increasingly large volume of medical images that need to be processed. It has been demonstrated that for medical images the transmission from local to clouds is much more expensive than the computation in the clouds itself. Towards this, 3D image compression techniques have been widely applied to reduce the data traffic. However, most of the existing image compression techniques are developed around human vision, i.e., they are designed to minimize distortions that can be perceived by human eyes. In this paper, we will use deep learning based medical image segmentation as a vehicle and demonstrate that interestingly, machine and human view the compression quality differently. Medical images compressed with good quality w.r.t. human vision may result in inferior segmentation accuracy. We then design a machine vision oriented 3D image compression framework tailored for segmentation using DNNs. Our method automatically extracts and retains ima

ge features that are most important to the segmentation. Comprehensive experimen ts on widely adopted segmentation frameworks with HVSMR 2016 challenge dataset s how that our method can achieve significantly higher segmentation accuracy at th e same compression rate, or much better compression rate under the same segmenta tion accuracy, when compared with the existing JPEG 2000 method. To the best of the authors' knowledge, this is the first machine vision guided medical image co mpression framework for segmentation in the clouds.
****

PointPillars: Fast Encoders for Object Detection From Point Clouds
Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, Oscar B eijbom; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Re cognition (CVPR), 2019, pp. 12697-12705
Object detection in point clouds is an important aspect of many robotics applica tions such as autonomous driving. In this paper, we consider the problem of enco ding a point cloud into a format appropriate for a downstream detection pipeline . Recent literature suggests two types of encoders; fixed encoders tend to be fa st but sacrifice accuracy, while encoders that are learned from data are more ac curate, but slower. In this work, we propose PointPillars, a novel encoder which utilizes PointNets to learn a representation of point clouds organized in verti cal columns (pillars). While the encoded features can be used with any standard 2D convolutional detection architecture, we further propose a lean downstream ne twork. Extensive experimentation shows that PointPillars outperforms previous en coders with respect to both speed and accuracy by a large margin. Despite only u sing lidar, our full detection pipeline significantly outperforms the state of t he art, even among fusion methods, with respect to both the 3D and bird's eye vi ew KITTI benchmarks. This detection performance is achieved while running at 62 Hz: a 2 - 4 fold runtime improvement. A faster version of our method matches the state of the art at 105 Hz. These benchmarks suggest that PointPillars is an ap propriate encoding for object detection in point clouds.
****

Motion Estimation of Non-Holonomic Ground Vehicles From a Single Feature Corresp ondence Measured Over N Views
Kun Huang, Yifu Wang, Laurent Kneip; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12706-12715
The planar motion of ground vehicles is often non-holonomic, which enables a sol ution of the two-view relative pose problem from a single point feature correspo ndence. Man-made environments such as underground parking lots are however domin ated by line features. Inspired by the planar tri-focal tensor and its ability t o handle lines, we establish an n-linear constraint on the locally circular moti on of non-holonomic vehicles able to handle an arbitrarily large and dense windo w of views. We prove that this stays a uni-variate problem under the assumption of locally constant vehicle speed, and it can transparently handle both point an d vertical line correspondences. In particular, we prove that an application of Viete's formulas for extrapolating trigonometric functions of angle multiples an d the Weierstrass substitution casts the problem as one that merely seeks the ro ots of a uni-variate polynomial. We present the complete theory of this novel so lver, and test it on both simulated and real data. Our results prove that it suc cessfully handles a variety of relevant scenarios, eventually outperforming the 1-point two-view solver.
****

From Coarse to Fine: Robust Hierarchical Localization at Large Scale
Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, Marcin Dymczyk; Proceedin gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12716-12725
Robust and accurate visual localization is a fundamental capability for numerous applications, such as autonomous driving, mobile robotics, or augmented reality . It remains, however, a challenging task, particularly for large-scale environm ents and in presence of significant appearance changes. State-of-the-art methods not only struggle with such scenarios, but are often too resource intensive for certain real-time applications. In this paper we propose HF-Net, a hierarchical

localization approach based on a monolithic CNN that simultaneously predicts local features and global descriptors for accurate 6-DoF localization. We exploit the coarse-to-fine localization paradigm: we first perform a global retrieval to obtain location hypotheses and only later match local features within those candidate places. This hierarchical approach incurs significant runtime savings and makes our system suitable for real-time operation. By leveraging learned descriptors, our method achieves remarkable localization robustness across large variations of appearance and sets a new state-of-the-art on two challenging benchmarks for large-scale localization.
****

Large Scale High-Resolution Land Cover Mapping With Multi-Resolution Data
Caleb Robinson, Le Hou, Kolya Malkin, Rachel Soobitsky, Jacob Czawlytko, Bistra Dilkina, Nebojsa Jojic; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12726-12735
In this paper we propose multi-resolution data fusion methods for deep learning-based high-resolution land cover mapping from aerial imagery. The land cover mapping problem, at country-level scales, is challenging for common deep learning methods due to the scarcity of high-resolution labels, as well as variation in geography and quality of input images. On the other hand, multiple satellite imagery and low-resolution ground truth label sources are widely available, and can be used to improve model training efforts. Our methods include: introducing low-resolution satellite data to smooth quality differences in high-resolution input, exploiting low-resolution labels with a dual loss function, and pairing scarce high-resolution labels with inputs from several points in time. We train models that are able to generalize from a portion of the Northeast United States, where we have high-resolution land cover labels, to the rest of the US. With these models, we produce the first high-resolution (1-meter) land cover map of the contiguous US, consisting of over 8 trillion pixels. We demonstrate the robustness and potential applications of this data in a case study with domain experts and develop a web application to share our results. This work is practically useful, and can be applied to other locations over the earth as high-resolution imagery becomes more widely available even as high-resolution labeled land cover data remains sparse.
****

Leveraging Heterogeneous Auxiliary Tasks to Assist Crowd Counting
Muming Zhao, Jian Zhang, Chongyang Zhang, Wenjun Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12736-12745
Crowd counting is a challenging task in the presence of drastic scale variations, the clutter background, and severe occlusions, etc. Existing CNN-based counting methods tackle these challenges mainly by fusing either multi-scale or multi-context features to generate robust representations. In this paper, we propose to address these issues by leveraging the heterogeneous attributes compounded in the density map. We identify three geometric/semantic/numeric attributes essentially important to the density estimation, and demonstrate how to effectively utilize these heterogeneous attributes to assist the crowd counting by formulating them into multiple auxiliary tasks. With the multi-fold regularization effects induced by the auxiliary tasks, the backbone CNN model is driven to embed desired properties explicitly and thus gains robust representations towards more accurate density estimation. Extensive experiments on three challenging crowd counting datasets have demonstrated the effectiveness of the proposed approach.
****