Feature Weighting and Boosting for Few-Shot Segmentation
Khoi Nguyen, Sinisa Todorovic; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 622-631
This paper is about few-shot segmentation of foreground objects in images. We train a CNN on small subsets of training images, each mimicking the few-shot setting. In each subset, one image serves as the query and the other(s) as support image(s) with ground-truth segmentation. The CNN first extracts feature maps from the query and support images. Then, a class feature vector is computed as an average of the support's feature maps over the known foreground. Finally, the target object is segmented in the query image by using a cosine similarity between the class feature vector and the query's feature map. We make two contributions by: (1) Improving discriminativeness of features so their activations are high on the foreground and low elsewhere; and (2) Boosting inference with an ensemble of experts guided with the gradient of loss incurred when segmenting the support images in testing. Our evaluations on the PASCAL-5i and COCO-20i datasets demonstrate that we significantly outperform existing approaches.
****

Image2StyleGAN: How to Embed Images Into the StyleGAN Latent Space?
Rameen Abdal, Yipeng Qin, Peter Wonka; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4432-4441
We propose an efficient algorithm to embed a given image into the latent space of StyleGAN. This embedding enables semantic image editing operations that can be applied to existing photographs. Taking the StyleGAN trained on the FFHD dataset as an example, we show results for image morphing, style transfer, and expression transfer. Studying the results of the embedding algorithm provides valuable insights into the structure of the StyleGAN latent space. We propose a set of experiments to test what class of images can be embedded, how they are embedded, what latent space is suitable for embedding, and if the embedding is semantically meaningful.
****

Controllable Artistic Text Style Transfer via Shape-Matching GAN
Shuai Yang, Zhangyang Wang, Zhaowen Wang, Ning Xu, Jiaying Liu, Zongming Guo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4442-4451
Artistic text style transfer is the task of migrating the style from a source image to the target text to create artistic typography. Recent style transfer methods have considered texture control to enhance usability. However, controlling the stylistic degree in terms of shape deformation remains an important open challenge. In this paper, we present the first text style transfer network that allows for real-time control of the crucial stylistic degree of the glyph through an adjustable parameter. Our key contribution is a novel bidirectional shape matching framework to establish an effective glyph-style mapping at various deformation levels without paired ground truth. Based on this idea, we propose a scale-controllable module to empower a single network to continuously characterize the multi-scale shape features of the style image and transfer these features to the target text. The proposed method demonstrates its superiority over previous state-of-the-arts in generating diverse, controllable and high-quality stylized text.
****

Understanding Generalized Whitening and Coloring Transform for Universal Style Transfer
Tai-Yin Chiu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4452-4460
Style transfer is a task of rendering images in the styles of other images. In the past few years, neural style transfer has achieved a great success in this task, yet suffers from either the inability to generalize to unseen style images or fast style transfer. Recently, an universal style transfer technique that applies zero-phase component analysis (ZCA) for whitening and coloring image features realizes fast and arbitrary style transfer. However, using ZCA for style transfer is empirical and does not have any theoretical support. In addition, other w

hitening and coloring transforms (WCT) than ZCA have not been investigated. In t
his report, we generalize ZCA to the general form of WCT, provide an analytical
performance analysis from the angle of neural style transfer, and show why ZCA i
s a good choice for style transfer among different WCTs and why some WCTs are no
t well applicable for style transfer.
****

Learning Implicit Generative Models by Matching Perceptual Features
Cicero Nogueira dos Santos, Youssef Mroueh, Inkit Padhi, Pierre Dognin; Proce
edings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019,
pp. 4461-4470
Perceptual features (PFs) have been used with great success in tasks such as tra
nsfer learning, style transfer, and super-resolution. However, the efficacy of P
Fs as key source of information for learning generative models is not well studi
ed. We investigate here the use of PFs in the context of learning implicit gener
ative models through moment matching (MM). More specifically, we propose a new e
ffective MM approach that learns implicit generative models by performing mean a
nd covariance matching of features extracted from pretrained ConvNets. Our propo
sed approach improves upon existing MM methods by: (1) breaking away from the pr
oblematic min/max game of adversarial learning; (2) avoiding online learning of
kernel functions; and (3) being efficient with respect to both number of used mo
ments and required minibatch size. Our experimental results demonstrate that, du
e to the expressiveness of PFs from pretrained deep ConvNets, our method achieve
s state-of-the-art results for challenging benchmarks.
****

Free-Form Image Inpainting With Gated Convolution
Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, Thomas S. Huang; Proc
eedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019
, pp. 4471-4480
We present a generative image inpainting system to complete images with free-for
m mask and guidance. The system is based on gated convolutions learned from mill
ions of images without additional labelling efforts. The proposed gated convolut
ion solves the issue of vanilla convolution that treats all input pixels as vali
d ones, generalizes partial convolution by providing a learnable dynamic feature
selection mechanism for each channel at each spatial location across all layers
. Moreover, as free-form masks may appear anywhere in images with any shape, glo
bal and local GANs designed for a single rectangular mask are not applicable. Th
us, we also present a patch-based GAN loss, named SN-PatchGAN, by applying spect
ral-normalized discriminator on dense image patches. SN-PatchGAN is simple in fo
rmulation, fast and stable in training. Results on automatic image inpainting an
d user-guided extension demonstrate that our system generates higher-quality and
more flexible results than previous methods. Our system helps user quickly remo
ve distracting objects, modify image layouts, clear watermarks and edit faces. C
ode, demo and models are available at: https://github.com/JiahuiYu/generative_in
painting.
****

FiNet: Compatible and Diverse Fashion Image Inpainting
Xintong Han, Zuxuan Wu, Weilin Huang, Matthew R. Scott, Larry S. Davis; Proc
eedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019
, pp. 4481-4491
Visual compatibility is critical for fashion analysis, yet is missing in existin
g fashion image synthesis systems. In this paper, we propose to explicitly model
visual compatibility through fashion image inpainting. We present Fashion Inpai
nting Networks (FiNet), a two-stage image-to-image generation framework that is
able to perform compatible and diverse inpainting. Disentangling the generation
of shape and appearance to ensure photorealistic results, our framework consists
of a shape generation network and an appearance generation network. More import
antly, for each generation network, we introduce two encoders interacting with o
ne another to learn latent codes in a shared compatibility space. The latent rep
resentations are jointly optimized with the corresponding generation network to
condition the synthesis process, encouraging a diverse set of generated results

that are visually compatible with existing fashion garments. In addition, our fr
amework is readily extended to clothing reconstruction and fashion transfer. Ext
ensive experiments on fashion synthesis quantitatively and qualitatively demonst
rate the effectiveness of our method.
****

InGAN: Capturing and Retargeting the "DNA" of a Natural Image
Assaf Shocher, Shai Bagon, Phillip Isola, Michal Irani; Proceedings of the IE
EE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4492-4501
Generative Adversarial Networks (GANs) typically learn a distribution of images
in a large image dataset, and are then able to generate new images from this dis
tribution. However, each natural image has its own internal statistics, captured
by its unique distribution of patches. In this paper we propose an "Internal GA
N" (InGAN) -- an image-specific GAN -- which trains on a single input image and
learns its internal distribution of patches. It is then able to synthesize a ple
thora of new natural images of significantly different sizes, shapes and aspect-
ratios - all with the same internal patch-distribution (same "DNA") as the input
image. In particular, despite large changes in global size/shape of the image,
all elements inside the image maintain their local size/shape. InGAN is fully un
supervised, requiring no additional data other than the input image itself. Once
trained on the input image, it can remap the input to any size or shape in a si
ngle feedforward pass, while preserving the same internal patch distribution. In
GAN provides a unified framework for a variety of tasks, bridging the gap betwee
n textures and natural images.
****

Seeing What a GAN Cannot Generate
David Bau, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bol
ei Zhou, Antonio Torralba; Proceedings of the IEEE/CVF International Conference
on Computer Vision (ICCV), 2019, pp. 4502-4511
Despite the success of Generative Adversarial Networks (GANs), mode collapse rem
ains a serious issue during GAN training. To date, little work has focused on un
derstanding and quantifying which modes have been dropped by a model. In this wo
rk, we visualize mode collapse at both the distribution level and the instance l
evel. First, we deploy a semantic segmentation network to compare the distributi
on of segmented objects in the generated images with the target distribution in
the training set. Differences in statistics reveal object classes that are omitt
ed by a GAN. Second, given the identified omitted object classes, we visualize t
he GAN's omissions directly. In particular, we compare specific differences betw
een individual photos and their approximate inversions by a GAN. To this end, we
relax the problem of inversion and solve the tractable problem of inverting a G
AN layer instead of the entire generator. Finally, we use this framework to anal
yze several recent GANs trained on multiple datasets and identify their typical
failure cases.
****

COCO-GAN: Generation by Parts via Conditional Coordinating
Chieh Hubert Lin, Chia-Che Chang, Yu-Sheng Chen, Da-Cheng Juan, Wei Wei, Hw
ann-Tzong Chen; Proceedings of the IEEE/CVF International Conference on Computer
Vision (ICCV), 2019, pp. 4512-4521
Humans can only interact with part of the surrounding environment due to biologi
cal restrictions. Therefore, we learn to reason the spatial relationships across
a series of observations to piece together the surrounding environment. Inspire
d by such behavior and the fact that machines also have computational constraint
s, we propose COnditional COordinate GAN (COCO-GAN) of which the generator gener
ates images by parts based on their spatial coordinates as the condition. On the
other hand, the discriminator learns to justify realism across multiple assembl
ed patches by global coherence, local appearance, and edge-crossing continuity.
Despite the full images are never manipulated during training, we show that COCO
-GAN can produce state-of-the-art-quality full images during inference. We furth
er demonstrate a variety of novel applications enabled by our coordinate-aware f
ramework. First, we perform extrapolation to the learned coordinate manifold and
generate off-the-boundary patches. Combining with the originally generated full

image, COCO-GAN can produce images that are larger than training samples, which we called "beyond-boundary generation". We then showcase panorama generation within a cylindrical coordinate system that inherently preserves horizontally cyclic topology. On the computation side, COCO-GAN has a built-in divide-and-conquer paradigm that reduces memory requisition during training and inference, provides high-parallelism, and can generate parts of images on-demand.
****

Neural Turtle Graphics for Modeling City Road Layouts
Hang Chu, Daiqing Li, David Acuna, Amlan Kar, Maria Shugrina, Xinkai Wei, Ming-Yu Liu, Antonio Torralba, Sanja Fidler; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4522-4530
We propose Neural Turtle Graphics (NTG), a novel generative model for spatial graphs, and demonstrate its applications in modeling city road layouts. Specifically, we represent the road layout using a graph where nodes in the graph represent control points and edges in the graph represents road segments. NTG is a sequential generative model parameterized by a neural network. It iteratively generates a new node and an edge connecting to an existing node conditioned on the current graph. We train NTG on Open Street Map data and show it outperforms existing approaches using a set of diverse performance metrics. Moreover, our method allows users to control styles of generated road layouts mimicking existing cities as well as to sketch a part of the city road layout to be synthesized. In addition to synthesis, the proposed NTG finds uses in an analytical task of aerial road parsing. Experimental results show that it achieves state-of-the-art performance on the SpaceNet dataset.
****

Texture Fields: Learning Texture Representations in Function Space
Michael Oechsle, Lars Mescheder, Michael Niemeyer, Thilo Strauss, Andreas Geiger; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4531-4540
In recent years, substantial progress has been achieved in learning-based reconstruction of 3D objects. At the same time, generative models were proposed that can generate highly realistic images. However, despite this success in these closely related tasks, texture reconstruction of 3D objects has received little attention from the research community and state-of-the-art methods are either limited to comparably low resolution or constrained experimental setups. A major reason for these limitations is that common representations of texture are inefficient or hard to interface for modern deep learning techniques. In this paper, we propose Texture Fields, a novel texture representation which is based on regressing a continuous 3D function parameterized with a neural network. Our approach circumvents limiting factors like shape discretization and parameterization, as the proposed texture representation is independent of the shape representation of the 3D object. We show that Texture Fields are able to represent high frequency texture and naturally blend with modern deep learning techniques. Experimentally, we find that Texture Fields compare favorably to state-of-the-art methods for conditional texture reconstruction of 3D objects and enable learning of probabilistic generative models for texturing unseen 3D models. We believe that Texture Fields will become an important building block for the next generation of generative 3D models.
****

PointFlow: 3D Point Cloud Generation With Continuous Normalizing Flows
Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, Bharath Hariharan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4541-4550
As 3D point clouds become the representation of choice for multiple vision and graphics applications, the ability to synthesize or reconstruct high-resolution, high-fidelity point clouds becomes crucial. Despite the recent success of deep learning models in discriminative tasks of point clouds, generating point clouds remains challenging. This paper proposes a principled probabilistic framework to generate 3D point clouds by modeling them as a distribution of distributions. Specifically, we learn a two-level hierarchy of distributions where the first lev

el is the distribution of shapes and the second level is the distribution of poi
nts given a shape. This formulation allows us to both sample shapes and sample a
n arbitrary number of points from a shape. Our generative model, named PointFlow
, learns each level of the distribution with a continuous normalizing flow. The
invertibility of normalizing flows enables the computation of the likelihood dur
ing training and allows us to train our model in the variational inference frame
work. Empirically, we demonstrate that PointFlow achieves state-of-the-art perfo
rmance in point cloud generation. We additionally show that our model can faithf
ully reconstruct point clouds and learn useful representations in an unsupervise
d manner. The code is available at https://github.com/stevenygd/PointFlow.
****

Meta-Sim: Learning to Generate Synthetic Datasets
Amlan Kar, Aayush Prakash, Ming-Yu Liu, Eric Cameracci, Justin Yuan, Matt R
usiniak, David Acuna, Antonio Torralba, Sanja Fidler; Proceedings of the IEEE
/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4551-4560
Training models to high-end performance requires availability of large labeled d
atasets, which are expensive to get. The goal of our work is to automatically sy
nthesize labeled datasets that are relevant for a downstream task. We propose Me
ta-Sim, which learns a generative model of synthetic scenes, and obtain images a
s well as its corresponding ground-truth via a graphics engine. We parametrize o
ur dataset generator with a neural network, which learns to modify attributes of
 scene graphs obtained from probabilistic scene grammars, so as to minimize the
distribution gap between its rendered outputs and target data. If the real datas
et comes with a small labeled validation set, we additionally aim to optimize a
meta-objective, i.e. downstream task performance. Experiments show that the prop
osed method can greatly improve content generation quality over a human-engineer
ed probabilistic scene grammar, both qualitatively and quantitatively as measure
d by performance on a downstream task.
****

Specifying Object Attributes and Relations in Interactive Scene Generation
Oron Ashual, Lior Wolf; Proceedings of the IEEE/CVF International Conference on
 Computer Vision (ICCV), 2019, pp. 4561-4569
We introduce a method for the generation of images from an input scene graph. Th
e method separates between a layout embedding and an appearance embedding. The d
ual embedding leads to generated images that better match the scene graph, have
higher visual quality, and support more complex scene graphs. In addition, the e
mbedding scheme supports multiple and diverse output images per scene graph, whi
ch can be further controlled by the user. We demonstrate two modes of per-object
 control: (i) importing elements from other images, and (ii) navigation in the o
bject space by selecting an appearance archetype. Our code is publicly available
 at https://www.github.com/ashual/scene_generation.
****

SinGAN: Learning a Generative Model From a Single Natural Image
Tamar Rott Shaham, Tali Dekel, Tomer Michaeli; Proceedings of the IEEE/CVF Int
ernational Conference on Computer Vision (ICCV), 2019, pp. 4570-4580
We introduce SinGAN, an unconditional generative model that can be learned from
a single natural image. Our model is trained to capture the internal distributio
n of patches within the image, and is then able to generate high quality, divers
e samples that carry the same visual content as the image. SinGAN contains a pyr
amid of fully convolutional GANs, each responsible for learning the patch distri
bution at a different scale of the image. This allows generating new samples of
arbitrary size and aspect ratio, that have significant variability, yet maintain
 both the global structure and the fine textures of the training image. In contr
ast to previous single image GAN schemes, our approach is not limited to texture
 images, and is not conditional (i.e. it generates samples from noise). User stu
dies confirm that the generated samples are commonly confused to be real images.
 We illustrate the utility of SinGAN in a wide range of image manipulation tasks
.
****

VaTeX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language R

esearch
Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, William Yang Wang ; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) , 2019, pp. 4581-4591
We present a new large-scale multilingual video description dataset, VATEX, whic h contains over 41,250 videos and 825,000 captions in both English and Chinese. Among the captions, there are over 206,000 English-Chinese parallel translation pairs. Compared to the widely-used MSR-VTT dataset, \vatex is multilingual, larg er, linguistically complex, and more diverse in terms of both video and natural language descriptions. We also introduce two tasks for video-and-language resear ch based on \vatex: (1) Multilingual Video Captioning, aimed at describing a vid eo in various languages with a compact unified captioning model, and (2) Video-g uided Machine Translation, to translate a source language description into the t arget language using the video information as additional spatiotemporal context. Extensive experiments on the \vatex dataset show that, first, the unified multi lingual model can not only produce both English and Chinese descriptions for a v ideo more efficiently, but also offer improved performance over the monolingual models. Furthermore, we demonstrate that the spatiotemporal video context can be effectively utilized to align source and target languages and thus assist machi ne translation. In the end, we discuss the potentials of using \vatex for other video-and-language research.
****
A Graph-Based Framework to Bridge Movies and Synopses
Yu Xiong, Qingqiu Huang, Lingfeng Guo, Hang Zhou, Bolei Zhou, Dahua Lin; Pr oceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 20 19, pp. 4592-4601
Inspired by the remarkable advances in video analytics, research teams are stepp ing towards a greater ambition - movie understanding. However, compared to those activity videos in conventional datasets, movies are significantly different. G enerally, movies are much longer and consist of much richer temporal structures. More importantly, the interactions among characters play a central role in expr essing the underlying story. To facilitate the efforts along this direction, we construct a dataset called Movie Synopses Associations (MSA) over 327 movies, wh ich provides a synopsis for each movie, together with annotated associations bet ween synopsis paragraphs and movie segments. On top of this dataset, we develop a framework to perform matching between movie segments and synopsis paragraphs. This framework integrates different aspects of a movie, including event dynamics and character interactions, and allows them to be matched with parsed paragraph s, based on a graph-based formulation. Our study shows that the proposed framewo rk remarkably improves the matching accuracy over conventional feature-based met hods. It also reveals the importance of narrative structures and character inter actions in movie understanding. Dataset and code are available at: https://ycxio ooong.github.io/projects/moviesyn
****
From Strings to Things: Knowledge-Enabled VQA Model That Can Read and Reason
Ajeet Kumar Singh, Anand Mishra, Shashank Shekhar, Anirban Chakraborty; Proce edings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4602-4612
Text present in images are not merely strings, they provide useful cues about th e image. Despite their utility in better image understanding, scene texts are no t used in traditional visual question answering (VQA) models. In this work, we p resent a VQA model which can read scene texts and perform reasoning on a knowled ge graph to arrive at an accurate answer. Our proposed model has three mutually interacting modules: i. proposal module to get word and visual content proposals from the image, ii. fusion module to fuse these proposals, question and knowled ge base to mine relevant facts, and represent these facts as multi-relational gr aph, iii. reasoning module to perform a novel gated graph neural network based r easoning on this graph. The performance of our knowledge-enabled VQA model is ev aluated on our newly introduced dataset, viz. text-KVQA. To the best of our know ledge, this is the first dataset which identifies the need for bridging text rec

ognition with knowledge graph based reasoning. Through extensive experiments, we show that our proposed method outperforms traditional VQA as well as question-answering over knowledge base-based methods on text-KVQA.
****

Counterfactual Critic Multi-Agent Training for Scene Graph Generation
Long Chen, Hanwang Zhang, Jun Xiao, Xiangnan He, Shiliang Pu, Shih-Fu Chang ; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) , 2019, pp. 4613-4623
Scene graphs --- objects as nodes and visual relationships as edges --- describe the whereabouts and interactions of objects in an image for comprehensive scene understanding. To generate coherent scene graphs, almost all existing methods exploit the fruitful visual context by modeling message passing among objects. For example, "person" on "bike" can help to determine the relationship "ride", which in turn contributes to the confidence of the two objects. However, we argue that the visual context is not properly learned by using the prevailing cross-entropy based supervised learning paradigm, which is not sensitive to graph inconsistency: errors at the hub or non-hub nodes should not be penalized equally. To this end, we propose a Counterfactual critic Multi-Agent Training (CMAT) approach. CMAT is a multi-agent policy gradient method that frames objects into cooperative agents, and then directly maximizes a graph-level metric as the reward. In particular, to assign the reward properly to each agent, CMAT uses a counterfactual baseline that disentangles the agent-specific reward by fixing the predictions of other agents. Extensive validations on the challenging Visual Genome benchmark show that CMAT achieves a state-of-the-art performance by significant gains under various settings and metrics.
****

Robust Change Captioning
Dong Huk Park, Trevor Darrell, Anna Rohrbach; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4624-4633
Describing what has changed in a scene can be useful to a user, but only if generated text focuses on what is semantically relevant. It is thus important to distinguish distractors (e.g. a viewpoint change) from relevant changes (e.g. an object has moved). We present a novel Dual Dynamic Attention Model (DUDA) to perform robust Change Captioning. Our model learns to distinguish distractors from semantic changes, localize the changes via Dual Attention over "before" and "after" images, and accurately describe them in natural language via Dynamic Speaker, by adaptively focusing on the necessary visual inputs (e.g. "before" or "after" image). To study the problem in depth, we collect a CLEVR-Change dataset, built off the CLEVR engine, with 5 types of scene changes. We benchmark a number of baselines on our dataset, and systematically study different change types and robustness to distractors. We show the superiority of our DUDA model in terms of both change captioning and localization. We also show that our approach is general, obtaining state-of-the-art results on the recent realistic Spot-the-Diff dataset which has no distractors.
****

Attention on Attention for Image Captioning
Lun Huang, Wenmin Wang, Jie Chen, Xiao-Yong Wei; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4634-4643
Attention mechanisms are widely used in current encoder/decoder frameworks of image captioning, where a weighted average on encoded vectors is generated at each time step to guide the caption decoding process. However, the decoder has little idea of whether or how well the attended vector and the given attention query are related, which could make the decoder give misled results. In this paper, we propose an Attention on Attention (AoA) module, which extends the conventional attention mechanisms to determine the relevance between attention results and queries. AoA first generates an information vector and an attention gate using the attention result and the current context, then adds another attention by applying element-wise multiplication to them and finally obtains the attended information, the expected useful knowledge. We apply AoA to both the encoder and the decoder of our image captioning model, which we name as AoA Network (AoANet). Exper

iments show that AoANet outperforms all previously published methods and achieves a new state-of-the-art performance of 129.8 CIDEr-D score on MS COCO Karpathy offline test split and 129.6 CIDEr-D (C40) score on the official online testing server. Code is available at https://github.com/husthuaan/AoANet.
****
Dynamic Graph Attention for Referring Expression Comprehension
Sibei Yang, Guanbin Li, Yizhou Yu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4644-4653
Referring expression comprehension aims to locate the object instance described by a natural language referring expression in an image. This task is compositional and inherently requires visual reasoning on top of the relationships among the objects in the image. Meanwhile, the visual reasoning process is guided by the linguistic structure of the referring expression. However, existing approaches treat the objects in isolation or only explore the first-order relationships between objects without being aligned with the potential complexity of the expression. Thus it is hard for them to adapt to the grounding of complex referring expressions. In this paper, we explore the problem of referring expression comprehension from the perspective of language-driven visual reasoning, and propose a dynamic graph attention network to perform multi-step reasoning by modeling both the relationships among the objects in the image and the linguistic structure of the expression. In particular, we construct a graph for the image with the nodes and edges corresponding to the objects and their relationships respectively, propose a differential analyzer to predict a language-guided visual reasoning process, and perform stepwise reasoning on top of the graph to update the compound object representation at every node. Experimental results demonstrate that the proposed method can not only significantly surpass all existing state-of-the-art algorithms across three common benchmark datasets, but also generate interpretable visual evidences for stepwise locating the objects referred to in complex language descriptions.
****
Visual Semantic Reasoning for Image-Text Matching
Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, Yun Fu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4654-4662
Image-text matching has been a hot research topic bridging the vision and language areas. It remains challenging because the current representation of image usually lacks global semantic concepts as in its corresponding text caption. To address this issue, we propose a simple and interpretable reasoning model to generate visual representation that captures key objects and semantic concepts of a scene. Specifically, we first build up connections between image regions and perform reasoning with Graph Convolutional Networks to generate features with semantic relationships. Then, we propose to use the gate and memory mechanism to perform global semantic reasoning on these relationship-enhanced features, select the discriminative information and gradually generate the representation for the whole scene. Experiments validate that our method achieves a new state-of-the-art for the image-text matching on MS-COCO and Flickr30K datasets. It outperforms the current best method by 6.8% relatively for image retrieval and 4.8% relatively for caption retrieval on MS-COCO (Recall@1 using 1K test set). On Flickr30K, our model improves image retrieval by 12.6% relatively and caption retrieval by 5.8% relatively (Recall@1).
****
Phrase Localization Without Paired Training Examples
Josiah Wang, Lucia Specia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4663-4672
Localizing phrases in images is an important part of image understanding and can be useful in many applications that require mappings between textual and visual information. Existing work attempts to learn these mappings from examples of phrase-image region correspondences (strong supervision) or from phrase-image pairs (weak supervision). We postulate that such paired annotations are unnecessary, and propose the first method for the phrase localization problem where neither training procedure nor paired, task-specific data is required. Our method is sim

ple but effective: we use off-the-shelf approaches to detect objects, scenes and colours in images, and explore different approaches to measure semantic similarity between the categories of detected visual elements and words in phrases. Experiments on two well-known phrase localization datasets show that this approach surpasses all weakly supervised methods by a large margin and performs very competitively to strongly supervised methods, and can thus be considered a strong baseline to the task. The non-paired nature of our method makes it applicable to any domain and where no paired phrase localization annotation is available.
****

Learning to Assemble Neural Module Tree Networks for Visual Grounding
Daqing Liu, Hanwang Zhang, Feng Wu, Zheng-Jun Zha; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4673-4682
Visual grounding, a task to ground (i.e., localize) natural language in images, essentially requires composite visual reasoning. However, existing methods over-simplify the composite nature of language into a monolithic sentence embedding or a coarse composition of subject-predicate-object triplet. In this paper, we propose to ground natural language in an intuitive, explainable, and composite fashion as it should be. In particular, we develop a novel modular network called Neural Module Tree network (NMTree) that regularizes the visual grounding along the dependency parsing tree of the sentence, where each node is a neural module that calculates visual attention according to its linguistic feature, and the grounding score is accumulated in a bottom-up direction where as needed. NMTree disentangles the visual grounding from the composite reasoning, allowing the former to only focus on primitive and easy-to-generalize patterns. To reduce the impact of parsing errors, we train the modules and their assembly end-to-end by using the Gumbel-Softmax approximation and its straight-through gradient estimator, accounting for the discrete nature of module assembly. Overall, the proposed NMTree consistently outperforms the state-of-the-arts on several benchmarks. Qualitative results show explainable grounding score calculation in great detail.
****

A Fast and Accurate One-Stage Approach to Visual Grounding
Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, Jiebo Luo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4683-4693
We propose a simple, fast, and accurate one-stage approach to visual grounding, inspired by the following insight. The performances of existing propose-and-rank two-stage methods are capped by the quality of the region candidates they propose in the first stage --- if none of the candidates could cover the ground truth region, there is no hope in the second stage to rank the right region to the top. To avoid this caveat, we propose a one-stage model that enables end-to-end joint optimization. The main idea is as straightforward as fusing a text query's embedding into the YOLOv3 object detector, augmented by spatial features so as to account for spatial mentions in the query. Despite being simple, this one-stage approach shows great potential in terms of both accuracy and speed for both phrase localization and referring expression comprehension, according to our experiments. Given these results along with careful investigations into some popular region proposals, we advocate for visual grounding a paradigm shift from the conventional two-stage methods to the one-stage framework.
****

Zero-Shot Grounding of Objects From Natural Language Queries
Arka Sadhu, Kan Chen, Ram Nevatia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4694-4703
A phrase grounding system localizes a particular object in an image referred to by a natural language query. In previous work, the phrases were restricted to have nouns that were encountered in training, we extend the task to Zero-Shot Grounding(ZSG) which can include novel, "unseen" nouns. Current phrase grounding systems use an explicit object detection network in a 2-stage framework where one stage generates sparse proposals and the other stage evaluates them. In the ZSG setting, generating appropriate proposals itself becomes an obstacle as the proposal generator is trained on the entities common in the detection and grounding d

atasets. We propose a new single-stage model called ZSGNet which combines the de
tector network and the grounding system and predicts classification scores and r
egression parameters. Evaluation of ZSG system brings additional subtleties due
to the influence of the relationship between the query and learned categories; w
e define four distinct conditions that incorporate different levels of difficult
y. We also introduce new datasets, sub-sampled from Flickr30k Entities and Visua
l Genome, that enable evaluations for the four conditions. Our experiments show
that ZSGNet achieves state-of-the-art performance on Flickr30k and ReferIt under
 the usual "seen" settings and performs significantly better than baseline in th
e zero-shot setting.
****

Towards Unconstrained End-to-End Text Spotting
Siyang Qin, Alessandro Bissacco, Michalis Raptis, Yasuhisa Fujii, Ying Xiao;
 Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),
 2019, pp. 4704-4714
We propose an end-to-end trainable network that can simultaneously detect and re
cognize text of arbitrary shape, making substantial progress on the open problem
 of reading scene text of irregular shape. We formulate arbitrary shape text det
ection as an instance segmentation problem; an attention model is then used to d
ecode the textual content of each irregularly shaped text region without rectifi
cation. To extract useful irregularly shaped text instance features from image s
cale features, we propose a simple yet effective RoI masking step. Additionally,
 we show that predictions from an existing multi-step OCR engine can be leverage
d as partially labeled training data, which leads to significant improvements in
 both the detection and recognition accuracy of our model. Our method surpasses
the state-of-the-art for end-to-end recognition tasks on the ICDAR15 (straight)
benchmark by 4.6%, and on the Total-Text (curved) benchmark by more than 16%.
****

What Is Wrong With Scene Text Recognition Model Comparisons? Dataset and Model A
nalysis
Jeonghun Baek, Geewook Kim, Junyeop Lee, Sungrae Park, Dongyoon Han, Sangdo
o Yun, Seong Joon Oh, Hwalsuk Lee; Proceedings of the IEEE/CVF International C
onference on Computer Vision (ICCV), 2019, pp. 4715-4723
Many new proposals for scene text recognition (STR) models have been introduced
in recent years. While each claim to have pushed the boundary of the technology,
 a holistic and fair comparison has been largely missing in the field due to the
 inconsistent choices of training and evaluation datasets. This paper addresses
this difficulty with three major contributions. First, we examine the inconsiste
ncies of training and evaluation datasets, and the performance gap results from
inconsistencies. Second, we introduce a unified four-stage STR framework that mo
st existing STR models fit into. Using this framework allows for the extensive e
valuation of previously proposed STR modules and the discovery of previously une
xplored module combinations. Third, we analyze the module-wise contributions to
performance in terms of accuracy, speed, and memory demand, under one consistent
 set of training and evaluation datasets. Such analyses clean up the hindrance o
n the current comparisons to understand the performance gain of the existing mod
ules. Our code is publicly available.
****

Sparse and Imperceivable Adversarial Attacks
Francesco Croce, Matthias Hein; Proceedings of the IEEE/CVF International Confe
rence on Computer Vision (ICCV), 2019, pp. 4724-4732
Neural networks have been proven to be vulnerable to a variety of adversarial at
tacks. From a safety perspective, highly sparse adversarial attacks are particul
arly dangerous. On the other hand the pixelwise perturbations of sparse attacks
are typically large and thus can be potentially detected. We propose a new black
-box technique to craft adversarial examples aiming at minimizing l_0-distance t
o the original image. Extensive experiments show that our attack is better or co
mpetitive to the state of the art. Moreover, we can integrate additional bounds
on the componentwise perturbation. Allowing pixels to change only in region of h
igh variation and avoiding changes along axis-aligned edges makes our adversaria

l examples almost non-perceivable. Moreover, we adapt the Projected Gradient Des
cent attack to the l_0-norm integrating componentwise constraints. This allows u
s to do adversarial training to enhance the robustness of classifiers against sp
arse and imperceivable adversarial manipulations.
****

Enhancing Adversarial Example Transferability With an Intermediate Level Attack
Qian Huang, Isay Katsman, Horace He, Zeqi Gu, Serge Belongie, Ser-Nam Lim;
Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),
2019, pp. 4733-4742
Neural networks are vulnerable to adversarial examples, malicious inputs crafted
 to fool trained models. Adversarial examples often exhibit black-box transfer,
meaning that adversarial examples for one model can fool another model. However,
 adversarial examples are typically overfit to exploit the particular architectu
re and feature representation of a source model, resulting in sub-optimal black-
box transfer attacks to other target models. We introduce the Intermediate Level
 Attack (ILA), which attempts to fine-tune an existing adversarial example for g
reater black-box transferability by increasing its perturbation on a pre-specifi
ed layer of the source model, improving upon state-of-the-art methods. We show t
hat we can select a layer of the source model to perturb without any knowledge o
f the target models while achieving high transferability. Additionally, we provi
de some explanatory insights regarding our method and the effect of optimizing f
or adversarial examples using intermediate feature maps.
****

Implicit Surface Representations As Layers in Neural Networks
Mateusz Michalkiewicz, Jhony K. Pontes, Dominic Jack, Mahsa Baktashmotlagh,
Anders Eriksson; Proceedings of the IEEE/CVF International Conference on Compute
r Vision (ICCV), 2019, pp. 4743-4752
Implicit shape representations, such as Level Sets, provide a very elegant formu
lation for performing computations involving curves and surfaces. However, inclu
ding implicit representations into canonical Neural Network formulations is far
from straightforward. This has consequently restricted existing approaches to sh
ape inference, to significantly less effective representations, perhaps most com
monly voxels occupancy maps or sparse point clouds. To overcome this limitation
we propose a novel formulation that permits the use of implicit representations
of curves and surfaces, of arbitrary topology, as individual layers in Neural Ne
twork architectures with end-to-end trainability. Specifically, we propose to re
present the output as an oriented level set of a continuous and discretised embe
dding function. We investigate the benefits of our approach on the task of 3D sh
ape prediction from a single image; and demonstrate its ability to produce a mor
e accurate reconstruction compared to voxel-based representations. We further sh
ow that our model is flexible and can be applied to a variety of shape inference
 problems.
****

A Tour of Convolutional Networks Guided by Linear Interpreters
Pablo Navarrete Michelini, Hanwen Liu, Yunhua Lu, Xingqun Jiang; Proceedings
of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 47
53-4762
Convolutional networks are large linear systems divided into layers and connecte
d by non-linear units. These units are the "articulations" that allow the networ
k to adapt to the input. To understand how a network manages to solve a problem
we must look at the articulated decisions in entirety. If we could capture the a
ctions of non-linear units for a particular input, we would be able to replay th
e whole system back and forth as if it was always linear. It would also reveal t
he actions of non-linearities because the resulting linear system, a Linear Inte
rpreter, depends on the input image. We introduce a hooking layer, called a Line
arScope, which allows us to run the network and the linear interpreter in parall
el. Its implementation is simple, flexible and efficient. From here we can make
many curious inquiries: how do these linear systems look like? When the rows and
 columns of the transformation matrix are images, how do they look like? What ty
pe of basis do these linear transformations rely on? The answers depend on the p

roblems presented, through which we take a tour to some popular architectures us ed for classification, super-resolution (SR) and image-to-image translation (I2I). For classification we observe that popular networks use a pixel-wise vote per class strategy and heavily rely on bias parameters. For SR and I2I we find that CNNs use wavelet-type basis similar to the human visual system. For I2I we reve al copy-move and template-creation strategies to generate outputs.
****

Small Steps and Giant Leaps: Minimal Newton Solvers for Deep Learning
Joao F. Henriques, Sebastien Ehrhardt, Samuel Albanie, Andrea Vedaldi; Procee dings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4763-4772
We propose a fast second-order method that can be used as a drop-in replacement for current deep learning solvers. Compared to stochastic gradient descent (SGD) , it only requires two additional forward-mode automatic differentiation operati ons per iteration, which has a computational cost comparable to two standard for ward passes and is easy to implement. Our method addresses long-standing issues with current second-order solvers, which invert an approximate Hessian matrix ev ery iteration exactly or by conjugate-gradient methods, procedures that are much slower than a SGD step. Instead, we propose to keep a single estimate of the gr adient projected by the inverse Hessian matrix, and update it once per iteration with just two passes over the network. This estimate has the same size and is s imilar to the momentum variable that is commonly used in SGD . No estimate of t he Hessian is maintained. We first validate our method, called CurveBall, on sma ll problems with known solutions (noisy Rosenbrock function and degenerate 2-lay er linear networks), where current deep learning solvers struggle. We then train several large models on CIFAR and ImageNet, including ResNet and VGG-f networks , where we demonstrate faster convergence with no hyperparameter tuning. We also show our optimiser's generality by testing on a large set of randomly generated architectures.
****

Semantic Adversarial Attacks: Parametric Transformations That Fool Deep Classifi ers
Ameya Joshi, Amitangshu Mukherjee, Soumik Sarkar, Chinmay Hegde; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 47 73-4783
Deep neural networks have been shown to exhibit an intriguing vulnerability to a dversarial input images corrupted with imperceptible perturbations. However, the majority of adversarial attacks assume global, fine-grained control over the im age pixel space. In this paper, we consider a different setting: what happens if the adversary could only alter specific attributes of the input image? These wo uld generate inputs that might be perceptibly different, but still natural-looki ng and enough to fool a classifier. We propose a novel approach to generate such "semantic" adversarial examples by optimizing a particular adversarial loss ove r the range-space of a parametric conditional generative model. We demonstrate i mplementations of our attacks on binary classifiers trained on face images, and show that such natural-looking semantic adversarial examples exist. We evaluate the effectiveness of our attack on synthetic and real data, and present detailed comparisons with existing attack methods. We supplement our empirical results w ith theoretical bounds that demonstrate the existence of such parametric adversa rial examples.
****

Hilbert-Based Generative Defense for Adversarial Examples
Yang Bai, Yan Feng, Yisen Wang, Tao Dai, Shu-Tao Xia, Yong Jiang; Proceedin gs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4784-4793
Adversarial perturbations of clean images are usually imperceptible for human ey es, but can confidently fool deep neural networks (DNNs) to make incorrect predi ctions. Such vulnerability of DNNs raises serious security concerns about their practicability in security-sensitive applications. To defend against such advers arial perturbations, recently developed PixelDefend purifies a perturbed image b

ased on PixelCNN in a raster scan order (row/column by row/column). However, suc
h scan mode insufficiently exploits the correlations between pixels, which furth
er limits its robustness performance. Therefore, we propose a more advanced Hilb
ert curve scan order to model the pixel dependencies in this paper. Hilbert curv
e could well preserve local consistency when mapping from 2-D image to 1-D vecto
r, thus the local features in neighboring pixels can be more effectively modeled
. Moreover, the defensive power can be further improved via ensembles of Hilbert
 curve with different orientations. Experimental results demonstrate the superio
rity of our method over the state-of-the-art defenses against various adversaria
l attacks.
****

On the Efficacy of Knowledge Distillation
Jang Hyun Cho, Bharath Hariharan; Proceedings of the IEEE/CVF International Con
ference on Computer Vision (ICCV), 2019, pp. 4794-4802
In this paper, we present a thorough evaluation of the efficacy of knowledge dis
tillation and its dependence on student and teacher architectures. Starting with
 the observation that more accurate teachers often don't make good teachers, we
attempt to tease apart the factors that affect knowledge distillation performanc
e. We find crucially that larger models do not often make better teachers. We sh
ow that this is a consequence of mismatched capacity, and that small students ar
e unable to mimic large teachers. We find typical ways of circumventing this (su
ch as performing a sequence of knowledge distillation steps) to be ineffective.
Finally, we show that this effect can be mitigated by stopping the teacher's tra
ining early. Our results generalize across datasets and models.
****

Sym-Parameterized Dynamic Inference for Mixed-Domain Image Translation
Simyung Chang, SeongUk Park, John Yang, Nojun Kwak; Proceedings of the IEEE/C
VF International Conference on Computer Vision (ICCV), 2019, pp. 4803-4811
Recent advances in image-to-image translation have led to some ways to generate
multiple domain images through a single network. However, there is still a limit
 in creating an image of a target domain without a dataset on it. We propose a m
ethod to expand the concept of `multi-domain' from data to the loss area, and to
 combine the characteristics of each domain to create an image. First, we introd
uce a sym-parameter and its learning method that can mix various losses and can
synchronize them with input conditions. Then, we propose Sym-parameterized Gener
ative Network (SGN) using it. Through experiments, we confirmed that SGN could m
ix the characteristics of various data and losses, and it is possible to transla
te images to any mixed-domain without ground truths, such as 30% Van Gogh and 20
% Monet and 40% snowy.
****

Better and Faster: Exponential Loss for Image Patch Matching
Shuang Wang, Yanfeng Li, Xuefeng Liang, Dou Quan, Bowu Yang, Shaowei Wei,
Licheng Jiao; Proceedings of the IEEE/CVF International Conference on Computer V
ision (ICCV), 2019, pp. 4812-4821
Recent studies on image patch matching are paying more attention on hard sample
learning, because easy samples do not contribute much to the network optimizatio
n. They have proposed various hard negative sample mining strategies, but very f
ew addressed this problem from the perspective of loss functions. Our research s
hows that the conventional Siamese and triplet losses treat all samples linearly
, thus make the training time consuming. Instead, we propose the exponential Sia
mese and triplet losses, which can naturally focus more on hard samples and put
less emphasis on easy ones, meanwhile, speed up the optimization. To assist the
exponential losses, we introduce the hard positive sample mining to further enha
nce the effectiveness. The extensive experiments demonstrate our proposal improv
es both metric and descriptor learning on several well accepted benchmarks, and
outperforms the state-of-the-arts on the UBC dataset. Moreover, it also shows a
better generalizability on cross-spectral image matching and image retrieval tas
ks.
****

Physical Adversarial Textures That Fool Visual Object Tracking

Rey Reza Wiyatno, Anqi Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4822-4831

We present a method for creating inconspicuous-looking textures that, when displayed as posters in the physical world, cause visual object tracking systems to become confused. As a target being visually tracked moves in front of such a poster, its adversarial texture makes the tracker lock onto it, thus allowing the target to evade. This adversarial attack evaluates several optimization strategies for fooling seldom-targeted regression models: non-targeted, targeted, and a newly-coined family of guided adversarial losses. Also, while we use the Expectation Over Transformation (EOT) algorithm to generate physical adversaries that fool tracking models when imaged under diverse conditions, we compare the impacts of different scene variables to find practical attack setups with high resulting adversarial strength and convergence speed. We further showcase that textures optimized using simulated scenes can confuse real-world tracking systems for cameras and robots.

****

Wasserstein GAN With Quadratic Transport Cost

Huidong Liu, Xianfeng Gu, Dimitris Samaras; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4832-4841

Wasserstein GANs are increasingly used in Computer Vision applications as they are easier to train. Previous WGAN variants mainly use the l_1 transport cost to compute the Wasserstein distance between the real and synthetic data distributions. The l_1 transport cost restricts the discriminator to be 1-Lipschitz. However, WGANs with l_1 transport cost were recently shown to not always converge. In this paper, we propose WGAN-QC, a WGAN with quadratic transport cost. Based on the quadratic transport cost, we propose an Optimal Transport Regularizer (OTR) to stabilize the training process of WGAN-QC. We prove that the objective of the discriminator during each generator update computes the exact quadratic Wasserstein distance between real and synthetic data distributions. We also prove that WGAN-QC converges to a local equilibrium point with finite discriminator updates per generator update. We show experimentally on a Dirac distribution that WGAN-QC converges, when many of the l_1 cost WGANs fail to [22]. Qualitative and quantitative results on the CelebA, CelebA-HQ, LSUN and the ImageNet dog datasets show that WGAN-QC is better than state-of-art GAN methods. WGAN-QC has much faster runtime than other WGAN variants.

****

Scalable Verified Training for Provably Robust Image Classification

Sven Gowal, Krishnamurthy (Dj) Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, Pushmeet Kohli; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4842-4851

Recent work has shown that it is possible to train deep neural networks that are provably robust to norm-bounded adversarial perturbations. Most of these methods are based on minimizing an upper bound on the worst-case loss over all possible adversarial perturbations. While these techniques show promise, they often result in difficult optimization procedures that remain hard to scale to larger networks. Through a comprehensive analysis, we show how a simple bounding technique, interval bound propagation (IBP), can be exploited to train large provably robust neural networks that beat the state-of-the-art in verified accuracy. While the upper bound computed by IBP can be quite weak for general networks, we demonstrate that an appropriate loss and clever hyper-parameter schedule allow the network to adapt such that the IBP bound is tight. This results in a fast and stable learning algorithm that outperforms more sophisticated methods and achieves state-of-the-art results on MNIST, CIFAR-10 and SVHN. It also allows us to train the largest model to be verified beyond vacuous bounds on a downscaled version of IMAGENET.

****

Differentiable Soft Quantization: Bridging Full-Precision and Low-Bit Neural Networks

Ruihao Gong, Xianglong Liu, Shenghu Jiang, Tianxiang Li, Peng Hu, Jiazhen L

in, Fengwei Yu, Junjie Yan; Proceedings of the IEEE/CVF International Conferen ce on Computer Vision (ICCV), 2019, pp. 4852-4861
Hardware-friendly network quantization (e.g., binary/uniform quantization) can e fficiently accelerate the inference and meanwhile reduce memory consumption of t he deep neural networks, which is crucial for model deployment on resource-limit ed devices like mobile phones. However, due to the discreteness of low-bit quant ization, existing quantization methods often face the unstable training process and severe performance degradation. To address this problem, in this paper we pr opose Differentiable Soft Quantization (DSQ) to bridge the gap between the full-precision and low-bit networks. DSQ can automatically evolve during training to gradually approximate the standard quantization. Owing to its differentiable pro perty, DSQ can help pursue the accurate gradients in backward propagation, and r educe the quantization loss in forward process with an appropriate clipping rang e. Extensive experiments over several popular network structures show that train ing low-bit neural networks with DSQ can consistently outperform state-of-the-ar t quantization methods. Besides, our first efficient implementation for deployin g 2 to 4-bit DSQ on devices with ARM architecture achieves up to 1.7x speed up, compared with the open-source 8-bit high-performance inference framework NCNN [3 1].
****

The LogBarrier Adversarial Attack: Making Effective Use of Decision Boundary Inf ormation
Chris Finlay, Aram-Alexandre Pooladian, Adam Oberman; Proceedings of the IEEE/ CVF International Conference on Computer Vision (ICCV), 2019, pp. 4862-4870
Adversarial attacks for image classification are small perturbations to images t hat are designed to cause misclassification by a model. Adversarial attacks form ally correspond to an optimization problem: find a minimum norm image perturbati on, constrained to cause misclassification. A number of effective attacks have b een developed. However, to date, no gradient-based attacks have used best practi ces from the optimization literature to solve this constrained minimization prob lem. We design a new untargeted attack, based on these best practices, using the well-regarded logarithmic barrier method. On average, our attack distance is si milar or better than all state-of-the-art attacks on benchmark datasets (MNIST, CIFAR10, ImageNet-1K). In addition, our method performs significantly better on the most challenging images, those which normally require larger perturbations f or misclassification. We employ the LogBarrier attack on several adversarially d efended models, and show that it adversarially perturbs all images more efficien tly than other attacks: the distance needed to perturb all images is significant ly smaller with the LogBarrier attack than with other state-of-the-art attacks.
****

Proximal Mean-Field for Neural Network Quantization
Thalaiyasingam Ajanthan, Puneet K. Dokania, Richard Hartley, Philip H. S. Tor r; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4871-4880
Compressing large Neural Networks (NN) by quantizing the parameters, while maint aining the performance is highly desirable due to reduced memory and time comple xity. In this work, we cast NN quantization as a discrete labelling problem, and by examining relaxations, we design an efficient iterative optimization procedu re that involves stochastic gradient descent followed by a projection. We prove that our simple projected gradient descent approach is, in fact, equivalent to a proximal version of the well-known mean-field method. These findings would allo w the decades-old and theoretically grounded research on MRF optimization to be used to design better network quantization schemes. Our experiments on standard classification datasets (MNIST, CIFAR10/100, TinyImageNet) with convolutional an d residual architectures show that our algorithm obtains fully-quantized network s with accuracies very close to the floating-point reference networks.
****

Improving Adversarial Robustness via Guided Complement Entropy
Hao-Yun Chen, Jhao-Hong Liang, Shih-Chieh Chang, Jia-Yu Pan, Yu-Ting Chen, Wei Wei, Da-Cheng Juan; Proceedings of the IEEE/CVF International Conference on

Computer Vision (ICCV), 2019, pp. 4881-4889
Adversarial robustness has emerged as an important topic in deep learning as carefully crafted attack samples can significantly disturb the performance of a model. Many recent methods have proposed to improve adversarial robustness by utilizing adversarial training or model distillation, which adds additional procedures to model training. In this paper, we propose a new training paradigm called Guided Complement Entropy (GCE) that is capable of achieving "adversarial defense for free," which involves no additional procedures in the process of improving adversarial robustness. In addition to maximizing model probabilities on the ground-truth class like cross-entropy, we neutralize its probabilities on the incorrect classes along with a "guided" term to balance between these two terms. We show in the experiments that our method achieves better model robustness with even better performance compared to the commonly used cross-entropy training objective. We also show that our method can be used orthogonal to adversarial training across well-known methods with noticeable robustness gain. To the best of our knowledge, our approach is the first one that improves model robustness without compromising performance.
****

A Geometry-Inspired Decision-Based Attack
Yujia Liu, Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4890-4898
Deep neural networks have recently achieved tremendous success in image classification. Recent studies have however shown that they are easily misled into incorrect classification decisions by adversarial examples. Adversaries can even craft attacks by querying the model in black-box settings, where no information about the model is released except its final decision. Such decision-based attacks usually require lots of queries, while real-world image recognition systems might actually restrict the number of queries. In this paper, we propose qFool, a novel decision-based attack algorithm that can generate adversarial examples using a small number of queries. The qFool method can drastically reduce the number of queries compared to previous decision-based attacks while reaching the same quality of adversarial examples. We also enhance our method by constraining adversarial perturbations in low-frequency subspace, which can make qFool even more computationally efficient. Altogether, we manage to fool commercial image recognition systems with a small number of queries, which demonstrates the actual effectiveness of our new algorithm in practice.
****

Universal Perturbation Attack Against Image Retrieval
Jie Li, Rongrong Ji, Hong Liu, Xiaopeng Hong, Yue Gao, Qi Tian; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4899-4908
Universal adversarial perturbations (UAPs), a.k.a. input-agnostic perturbations, has been proved to exist and be able to fool cutting-edge deep learning models on most of the data samples. Existing UAP methods mainly focus on attacking image classification models. Nevertheless, little attention has been paid to attacking image retrieval systems. In this paper, we make the first attempt in attacking image retrieval systems. Concretely, image retrieval attack is to make the retrieval system return irrelevant images to the query at the top ranking list. It plays an important role to corrupt the neighbourhood relationships among features in image retrieval attack. To this end, we propose a novel method to generate retrieval-against UAP to break the neighbourhood relationships of image features via degrading the corresponding ranking metric. To expand the attack method to scenarios with varying input sizes or untouchable network parameters, a multi-scale random resizing scheme and a ranking distillation strategy are proposed. We evaluate the proposed method on four widely-used image retrieval datasets, and report a significant performance drop in terms of different metrics, such as mAP and mP@10. Finally, we test our attack methods on the real-world visual search engine, i.e., Google Images, which demonstrates the practical potentials of our methods.
****

Bayesian Optimized 1-Bit CNNs
Jiaxin Gu, Junhe Zhao, Xiaolong Jiang, Baochang Zhang, Jianzhuang Liu, Guodong Guo, Rongrong Ji; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4909-4917
Deep convolutional neural networks (DCNNs) have dominated the recent developments in computer vision through making various record-breaking models. However, it is still a great challenge to achieve powerful DCNNs in resource-limited environments, such as on embedded devices and smart phones. Researchers have realized that 1-bit CNNs can be one feasible solution to resolve the issue; however, they are baffled by the inferior performance compared to the full-precision DCNNs. In this paper, we propose a novel approach, called Bayesian optimized 1-bit CNNs (denoted as BONNs), taking the advantage of Bayesian learning, a well-established strategy for hard problems, to significantly improve the performance of extreme 1-bit CNNs. We incorporate the prior distributions of full-precision kernels and features into the Bayesian framework to construct 1-bit CNNs in an end-to-end manner, which have not been considered in any previous related methods. The Bayesian losses are achieved with a theoretical support to optimize the network simultaneously in both continuous and discrete spaces, aggregating different losses jointly to improve the model capacity. Extensive experiments on the ImageNet and CIFAR datasets show that BONNs achieve the best classification performance compared to state-of-the-art 1-bit CNNs.
****

Rethinking ImageNet Pre-Training
Kaiming He, Ross Girshick, Piotr Dollar; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4918-4927
We report competitive results on object detection and instance segmentation on the COCO dataset using standard models trained from random initialization. The results are no worse than their ImageNet pre-training counterparts even when using the hyper-parameters of the baseline system (Mask R-CNN) that were optimized for fine-tuning pre-trained models, with the sole exception of increasing the number of training iterations so the randomly initialized models may converge. Training from random initialization is surprisingly robust; our results hold even when: (i) using only 10% of the training data, (ii) for deeper and wider models, and (iii) for multiple tasks and metrics. Experiments show that ImageNet pre-training speeds up convergence early in training, but does not necessarily provide regularization or improve final target task accuracy. To push the envelope we demonstrate 50.9 AP on COCO object detection without using any external data---a result on par with the top COCO 2017 competition results that used ImageNet pre-training. These observations challenge the conventional wisdom of ImageNet pre-training for dependent tasks and we expect these discoveries will encourage people to rethink the current de facto paradigm of `pre-training and fine-tuning' in computer vision.
****

Defending Against Universal Perturbations With Shared Adversarial Training
Chaithanya Kumar Mummadi, Thomas Brox, Jan Hendrik Metzen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4928-4937
Classifiers such as deep neural networks have been shown to be vulnerable against adversarial perturbations on problems with high-dimensional input space. While adversarial training improves the robustness of image classifiers against such adversarial perturbations, it leaves them sensitive to perturbations on a non-negligible fraction of the inputs. In this work, we show that adversarial training is more effective in preventing universal perturbations, where the same perturbation needs to fool a classifier on many inputs. Moreover, we investigate the trade-off between robustness against universal perturbations and performance on unperturbed data and propose an extension of adversarial training that handles this trade-off more gracefully. We present results for image classification and semantic segmentation to showcase that universal perturbations that fool a model hardened with adversarial training become clearly perceptible and show patterns of the target scene.
****

Adaptive Activation Thresholding: Dynamic Routing Type Behavior for Interpretability in Convolutional Neural Networks
Yiyou Sun, Sathya N. Ravi, Vikas Singh; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4938-4947
There is a growing interest in strategies that can help us understand or interpret neural networks -- that is, not merely provide a prediction, but also offer additional context explaining why and how. While many current methods offer tools to perform this analysis for a given (trained) network post-hoc, recent results (especially on capsule networks) suggest that when classes map to a few high level "concepts" in the preceding layers of the network, the behavior of the network is easier to interpret or explain. Such training may be accomplished via dynamic/EM routing where the network "routes" for individual classes (or subsets of images) are dynamic and involve few nodes even if the full network may not be sparse. In this paper, we show how a simple modification of the SGD scheme can help provide dynamic/EM routing type behavior in convolutional neural networks. Through extensive experiments, we evaluate the effect of this idea for interpretability where we obtain promising results, while also showing that no compromise in attainable accuracy is involved. Further, we show that the minor modification is seemingly ad-hoc, the new algorithm can be analyzed by an approximate method which provably matches known rates for SGD.
****

XRAI: Better Attributions Through Regions
Andrei Kapishnikov, Tolga Bolukbasi, Fernanda Viegas, Michael Terry; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4948-4957
Saliency methods can aid understanding of deep neural networks. Recent years have witnessed many improvements to saliency methods, as well as new ways for evaluating them. In this paper, we 1) present a novel region-based attribution method, XRAI, that builds upon integrated gradients (Sundararajan et al. 2017), 2) introduce evaluation methods for empirically assessing the quality of image-based saliency maps (Performance Information Curves (PICs)), and 3) contribute an axiom-based sanity check for attribution methods. Through empirical experiments and example results, we show that XRAI produces better results than other saliency methods for common models and the ImageNet dataset.
****

Guessing Smart: Biased Sampling for Efficient Black-Box Adversarial Attacks
Thomas Brunner, Frederik Diehl, Michael Truong Le, Alois Knoll; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4958-4966
We consider adversarial examples for image classification in the black-box decision-based setting. Here, an attacker cannot access confidence scores, but only the final label. Most attacks for this scenario are either unreliable or inefficient. Focusing on the latter, we show that a specific class of attacks, Boundary Attacks, can be reinterpreted as a biased sampling framework that gains efficiency from domain knowledge. We identify three such biases, image frequency, regional masks and surrogate gradients, and evaluate their performance against an ImageNet classifier. We show that the combination of these biases outperforms the state of the art by a wide margin. We also showcase an efficient way to attack the Google Cloud Vision API, where we craft convincing perturbations with just a few hundred queries. Finally, the methods we propose have also been found to work very well against strong defenses: Our targeted attack won second place in the NeurIPS 2018 Adversarial Vision Challenge.
****

Mask-Guided Attention Network for Occluded Pedestrian Detection
Yanwei Pang, Jin Xie, Muhammad Haris Khan, Rao Muhammad Anwer, Fahad Shahbaz Khan, Ling Shao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4967-4975
Pedestrian detection relying on deep convolution neural networks has made significant progress. Though promising results have been achieved on standard pedestrians, the performance on heavily occluded pedestrians remains far from satisfacto

ry. The main culprits are intra-class occlusions involving other pedestrians and inter-class occlusions caused by other objects, such as cars and bicycles. Thes e results in a multitude of occlusion patterns. We propose an approach for occlu ded pedestrian detection with the following contributions. First, we introduce a novel mask-guided attention network that fits naturally into popular pedestrian detection pipelines. Our attention network emphasizes on visible pedestrian reg ions while suppressing the occluded ones by modulating full body features. Secon d, we empirically demonstrate that coarse-level segmentation annotations provide reasonable approximation to their dense pixel-wise counterparts. Experiments ar e performed on CityPersons and Caltech datasets. Our approach sets a new state-o f-the-art on both datasets. Our approach obtains an absolute gain of 9.5% in log -average miss rate, compared to the best reported results [32] on the heavily oc cluded HO pedestrian set of CityPersons test set. Further, on the HO pedestrian set of Caltech dataset, our method achieves an absolute gain of 5.0% in log-aver age miss rate, compared to the best reported results [13]. Code and models are a vailable at: https://github.com/Leotju/MGAN.

****

Spectral Feature Transformation for Person Re-Identification

Chuanchen Luo, Yuntao Chen, Naiyan Wang, Zhaoxiang Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4976-4985

With the surge of deep learning techniques, the field of person re-identificatio n has witnessed rapid progress in recent years. Deep learning based methods focu s on learning a discriminative feature space where data points are clustered com pactly according to their corresponding identities. Most existing methods proces s data points individually or only involves a fraction of samples while building a similarity structure. They ignore dense informative connections among samples more or less. The lack of holistic observation eventually leads to inferior per formance. To relieve the issue, we propose to formulate the whole data batch as a similarity graph. Inspired by spectral clustering, a novel module termed Spect ral Feature Transformation is developed to facilitate the optimization of group-wise similarities. It adds no burden to the inference and can be applied to vari ous scenarios. As a natural extension, we further derive a lightweight re-rankin g method named Local Blurring Re-ranking which makes the underlying clustering s tructure around the probe set more compact. Empirical studies on four public ben chmarks show the superiority of the proposed method. Code is available at https: //github.com/LuckyDC/SFT_REID.

****

Permutation-Invariant Feature Restructuring for Correlation-Aware Image Set-Base d Recognition

Xiaofeng Liu, Zhenhua Guo, Site Li, Lingsheng Kong, Ping Jia, Jane You, B. V.K. Vijaya Kumar; Proceedings of the IEEE/CVF International Conference on Compu ter Vision (ICCV), 2019, pp. 4986-4996

We consider the problem of comparing the similarity of image sets with variable-quantity, quality and un-ordered heterogeneous images. We use feature restructur ing to exploit the correlations of both inner&inter-set images. Specifically, th e residual self-attention can effectively restructure the features using the oth er features within a set to emphasize the discriminative images and eliminate th e redundancy. Then, a sparse/collaborative learning-based dependency-guided repr esentation scheme reconstructs the probe features conditional to the gallery fea tures in order to adaptively align the two sets. This enables our framework to b e compatible with both verification and open-set identification. We show that th e parametric self-attention network and non-parametric dictionary learning can b e trained end-to-end by a unified alternative optimization scheme, and that the full framework is permutation-invariant. In the numerical experiments we conduct ed, our method achieves top performance on competitive image set/video-based fac e recognition and person re-identification benchmarks.

****

Improving Pedestrian Attribute Recognition With Weakly-Supervised Multi-Scale At tribute-Specific Localization

Chufeng Tang, Lu Sheng, Zhaoxiang Zhang, Xiaolin Hu; Proceedings of the IEEE/

CVF International Conference on Computer Vision (ICCV), 2019, pp. 4997-5006
Pedestrian attribute recognition has been an emerging research topic in the area of video surveillance. To predict the existence of a particular attribute, it is demanded to localize the regions related to the attribute. However, in this task, the region annotations are not available. How to carve out these attribute-related regions remains challenging. Existing methods applied attribute-agnostic visual attention or heuristic body-part localization mechanisms to enhance the local feature representations, while neglecting to employ attributes to define local feature areas. We propose a flexible Attribute Localization Module (ALM) to adaptively discover the most discriminative regions and learns the regional features for each attribute at multiple levels. Moreover, a feature pyramid architecture is also introduced to enhance the attribute-specific localization at low-levels with high-level semantic guidance. The proposed framework does not require additional region annotations and can be trained end-to-end with multi-level deep supervision. Extensive experiments show that the proposed method achieves state-of-the-art results on three pedestrian attribute datasets, including PETA, RAP, and PA-100K.
****

Correlation Congruence for Knowledge Distillation
Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, Zhaoning Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5007-5016
Most teacher-student frameworks based on knowledge distillation (KD) depend on a strong congruent constraint on instance level. However, they usually ignore the correlation between multiple instances, which is also valuable for knowledge transfer. In this work, we propose a new framework named correlation congruence for knowledge distillation (CCKD), which transfers not only the instance-level information but also the correlation between instances. Furthermore, a generalized kernel method based on Taylor series expansion is proposed to better capture the correlation between instances. Empirical experiments and ablation studies on image classification tasks (including CIFAR-100, ImageNet-1K) and metric learning tasks (including ReID and Face Recognition) show that the proposed CCKD substantially outperforms the original KD and other SOTA KD-based methods. The CCKD can be easily deployed in the majority of the teacher-student framework such as KD and hint-based learning methods.
****

Dynamic Curriculum Learning for Imbalanced Data Classification
Yiru Wang, Weihao Gan, Jie Yang, Wei Wu, Junjie Yan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5017-5026
Human attribute analysis is a challenging task in the field of computer vision. One of the significant difficulties is brought from largely imbalance-distributed data. Conventional techniques such as re-sampling and cost-sensitive learning require prior-knowledge to train the system. To address this problem, we propose a unified framework called Dynamic Curriculum Learning (DCL) to adaptively adjust the sampling strategy and loss weight in each batch, which results in better ability of generalization and discrimination. Inspired by curriculum learning, DCL consists of two-level curriculum schedulers: (1) sampling scheduler which manages the data distribution not only from imbalance to balance but also from easy to hard; (2) loss scheduler which controls the learning importance between classification and metric learning loss. With these two schedulers, we achieve state-of-the-art performance on the widely used face attribute dataset CelebA and pedestrian attribute dataset RAP.
****

Video Face Clustering With Unknown Number of Clusters
Makarand Tapaswi, Marc T. Law, Sanja Fidler; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5027-5036
Understanding videos such as TV series and movies requires analyzing who the characters are and what they are doing. We address the challenging problem of clustering face tracks based on their identity. Different from previous work in this area, we choose to operate in a realistic and difficult setting where: (i) the n

umber of characters is not known a priori; and (ii) face tracks belonging to min
or or background characters are not discarded. To this end, we propose Ball Clus
ter Learning (BCL), a supervised approach to carve the embedding space into ball
s of equal size, one for each cluster. The learned ball radius is easily transla
ted to a stopping criterion for iterative merging algorithms. This gives BCL the
 ability to estimate the number of clusters as well as their assignment, achievi
ng promising results on commonly used datasets. We also present a thorough discu
ssion of how existing metric learning literature can be adapted for this task.
****

Targeted Mismatch Adversarial Attack: Query With a Flower to Retrieve the Tower
Giorgos Tolias, Filip Radenovic, Ondrej Chum; Proceedings of the IEEE/CVF Inte
rnational Conference on Computer Vision (ICCV), 2019, pp. 5037-5046
Access to online visual search engines implies sharing of private user content -
- the query images. We introduce the concept of targeted mismatch attack for dee
p learning based retrieval systems to generate an adversarial image to conceal t
he query image. The generated image looks nothing like the user intended query,
but leads to identical or very similar retrieval results. Transferring attacks t
o fully unseen networks is challenging. We show successful attacks to partially
unknown systems, by designing various loss functions for the adversarial image c
onstruction. These include loss functions, for example, for unknown global pooli
ng operation or unknown input resolution by the retrieval system. We evaluate th
e attacks on standard retrieval benchmarks and compare the results retrieved wit
h the original and adversarial image.
****

Fashion++: Minimal Edits for Outfit Improvement
Wei-Lin Hsiao, Isay Katsman, Chao-Yuan Wu, Devi Parikh, Kristen Grauman; Pro
ceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 201
9, pp. 5047-5056
Given an outfit, what small changes would most improve its fashionability? This
question presents an intriguing new computer vision challenge. We introduce Fash
ion++, an approach that proposes minimal adjustments to a full-body clothing out
fit that will have maximal impact on its fashionability. Our model consists of a
 deep image generation neural network that learns to synthesize clothing conditi
oned on learned per-garment encodings. The latent encodings are explicitly facto
rized according to shape and texture, thereby allowing direct edits for both fit
/presentation and color/patterns/material, respectively. We show how to bootstra
p Web photos to automatically train a fashionability model, and develop an activ
ation maximization-style approach to transform the input image into its more fas
hionable self. The edits suggested range from swapping in a new garment to tweak
ing its color, how it is worn (e.g., rolling up sleeves), or its fit (e.g., maki
ng pants baggier). Experiments demonstrate that Fashion++ provides successful ed
its, both according to automated metrics and human opinion.
****

Semi-Supervised Pedestrian Instance Synthesis and Detection With Mutual Reinforc
ement
Si Wu, Sihao Lin, Wenhao Wu, Mohamed Azzam, Hau-San Wong; Proceedings of the
 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5057-506
6
We propose a GAN-based scene-specific instance synthesis and classification mode
l for semi-supervised pedestrian detection. Instead of collecting unreliable det
ections from unlabeled data, we adopt a class-conditional GAN for synthesizing p
edestrian instances to alleviate the problem of insufficient labeled data. With
the help of a base detector, we integrate pedestrian instance synthesis and dete
ction by including a post-refinement classifier (PRC) into a minimax game. A gen
erator and the PRC can mutually reinforce each other by synthesizing high-fideli
ty pedestrian instances and providing more accurate categorical information. Bot
h of them compete with a class-conditional discriminator and a class-specific di
scriminator, such that the four fundamental networks in our model can be jointly
 trained. In our experiments, we validate that the proposed model significantly
improves the performance of the base detector and achieves state-of-the-art resu

lts on multiple benchmarks. As shown in Figure 1, the result indicates the possi
bility of using inexpensively synthesized instances for improving semi-supervise
d detection models.
****

SILCO: Show a Few Images, Localize the Common Object
Tao Hu, Pascal Mettes, Jia-Hong Huang, Cees G. M. Snoek; Proceedings of the I
EEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5067-5076
Few-shot learning is a nascent research topic, motivated by the fact that tradit
ional deep learning requires tremendous amounts of data. In this work, we propos
e a new task along this research direction, we call few-shot common-localization
. Given a few weakly-supervised support images, we aim to localize the common ob
ject in the query image without any box annotation. This task differs from stand
ard few-shot settings, since we aim to address the localization problem, rather
than the global classification problem. To tackle this new problem, we propose a
network that aims to get the most out of the support and query images. To that
end, we introduce a spatial similarity module that searches the spatial commonal
ity among the given images. We furthermore introduce a feature reweighting modul
e to balance the influence of different support images through graph convolution
al networks. To evaluate few-shot common-localization, we repurpose and reorgani
ze the well-known Pascal VOC and MS-COCO datasets, as well as a video dataset fr
om ImageNet VID. Experiments on the new settings for few-shot common-localizatio
n shows the importance of searching for spatial similarity and feature reweighti
ng, outperforming baselines from related tasks.
****

A Deep Step Pattern Representation for Multimodal Retinal Image Registration
Jimmy Addison Lee, Peng Liu, Jun Cheng, Huazhu Fu; Proceedings of the IEEE/CV
F International Conference on Computer Vision (ICCV), 2019, pp. 5077-5086
This paper presents a novel feature-based method that is built upon a convolutio
nal neural network (CNN) to learn the deep representation for multimodal retinal
image registration. We coined the algorithm deep step patterns, in short DeepSP
a. Most existing deep learning based methods require a set of manually labeled t
raining data with known corresponding spatial transformations, which limits the
size of training datasets. By contrast, our method is fully automatic and scale
well to different image modalities with no human intervention. We generate featu
re classes from simple step patterns within patches of connecting edges formed b
y vascular junctions in multiple retinal imaging modalities. We leverage CNN to
learn and optimize the input patches to be used for image registration. Spatial
transformations are estimated based on the output possibility of the fully conne
cted layer of CNN for a pair of images. One of the key advantages of the propose
d algorithm is its robustness to non-linear intensity changes, which widely exis
t on retinal images due to the difference of acquisition modalities. We validate
our algorithm on extensive challenging datasets comprising poor quality multimo
dal retinal images which are adversely affected by pathologies (diseases), speck
le noise and low resolutions. The experimental results demonstrate the robustnes
s and accuracy over state-of-the-art multimodal image registration algorithms.
****

Deep Graphical Feature Learning for the Feature Matching Problem
Zhen Zhang, Wee Sun Lee; Proceedings of the IEEE/CVF International Conference o
n Computer Vision (ICCV), 2019, pp. 5087-5096
The feature matching problem is a fundamental problem in various areas of comput
er vision including image registration, tracking and motion analysis. Rich local
representation is a key part of efficient feature matching methods. However, wh
en the local features are limited to the coordinate of key points, it becomes ch
allenging to extract rich local representations. Traditional approaches use pair
wise or higher order handcrafted geometric features to get robust matching; this
requires solving NP-hard assignment problems. In this paper, we address this pr
oblem by proposing a graph neural network model to transform coordinates of feat
ure points into local features. With our local features, the traditional NP-hard
assignment problems are replaced with a simple assignment problem which can be
solved efficiently. Promising results on both synthetic and real datasets demons

trate the effectiveness of the proposed method.
****

Minimum Delay Object Detection From Video
Dong Lao, Ganesh Sundaramoorthi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5097-5106
We consider the problem of detecting objects, as they come into view, from videos in an online fashion. We provide the first real-time solution that is guaranteed to minimize the delay, i.e., the time between when the object comes in view and the declared detection time, subject to acceptable levels of detection accuracy. The method leverages modern CNN-based object detectors that operate on a single frame, to aggregate detection results over frames to provide reliable detection at a rate, specified by the user, in guaranteed minimal delay. To do this, we formulate the problem as a Quickest Detection problem, which provides the aforementioned guarantees. We derive our algorithms from this theory. We show in experiments, that with an overhead of just 50 fps, we can increase the number of correct detections and decrease the overall computational cost compared to running a modern single-frame detector.
****

Learning With Average Precision: Training Image Retrieval With a Listwise Loss
Jerome Revaud, Jon Almazan, Rafael S. Rezende, Cesar Roberto de Souza; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5107-5116
Image retrieval can be formulated as a ranking problem where the goal is to order database images by decreasing similarity to the query. Recent deep models for image retrieval have outperformed traditional methods by leveraging ranking-tailored loss functions, but important theoretical and practical problems remain. First, rather than directly optimizing the global ranking, they minimize an upper-bound on the essential loss, which does not necessarily result in an optimal mean average precision (mAP). Second, these methods require significant engineering efforts to work well, e.g., special pre-training and hard-negative mining. In this paper we propose instead to directly optimize the global mAP by leveraging recent advances in listwise loss formulations. Using a histogram binning approximation, the AP can be differentiated and thus employed to end-to-end learning. Compared to existing losses, the proposed method considers thousands of images simultaneously at each iteration and eliminates the need for ad hoc tricks. It also establishes a new state of the art on many standard retrieval benchmarks. Models and evaluation scripts have been made available at: https://europe.naverlabs.com/Deep-Image-Retrieval/.
****

Learning to Find Common Objects Across Few Image Collections
Amirreza Shaban, Amir Rahimi, Shray Bansal, Stephen Gould, Byron Boots, Richard Hartley; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5117-5126
Given a collection of bags where each bag is a set of images, our goal is to select one image from each bag such that the selected images are from the same object class. We model the selection as an energy minimization problem with unary and pairwise potential functions. Inspired by recent few-shot learning algorithms, we propose an approach to learn the potential functions directly from the data. Furthermore, we propose a fast greedy inference algorithm for energy minimization. We evaluate our approach on few-shot common object recognition as well as object co-localization tasks. Our experiments show that learning the pairwise and unary terms greatly improves the performance of the model over several well-known methods for these tasks. The proposed greedy optimization algorithm achieves performance comparable to state-of-the-art structured inference algorithms while being 10 times faster.
****

Weakly Aligned Cross-Modal Learning for Multispectral Pedestrian Detection
Lu Zhang, Xiangyu Zhu, Xiangyu Chen, Xu Yang, Zhen Lei, Zhiyong Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5127-5137

Multispectral pedestrian detection has shown great advantages under poor illumination conditions, since the thermal modality provides complementary information for the color image. However, real multispectral data suffers from the position shift problem, i.e. the color-thermal image pairs are not strictly aligned, making one object has different positions in different modalities. In deep learning based methods, this problem makes it difficult to fuse the feature maps from both modalities and puzzles the CNN training. In this paper, we propose a novel Aligned Region CNN (AR-CNN) to handle the weakly aligned multispectral data in an end-to-end way. Firstly, we design a Region Feature Alignment (RFA) module to capture the position shift and adaptively align the region features of the two modalities. Secondly, we present a new multimodal fusion method, which performs feature re-weighting to select more reliable features and suppress the useless ones. Besides, we propose a novel RoI jitter strategy to improve the robustness to unexpected shift patterns of different devices and system settings. Finally, since our method depends on a new kind of labelling: bounding boxes that match each modality, we manually relabel the KAIST dataset by locating bounding boxes in both modalities and building their relationships, providing a new KAIST-Paired Annotation. Extensive experimental validations on existing datasets are performed, demonstrating the effectiveness and robustness of the proposed method. Code and data are available at https://github.com/luzhang16/AR-CNN.
****

Deep Self-Learning From Noisy Labels
Jiangfan Han, Ping Luo, Xiaogang Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5138-5147
ConvNets achieve good results when training from clean data, but learning from noisy labels significantly degrades performances and remains challenging. Unlike previous works constrained by many conditions, making them infeasible to real noisy cases, this work presents a novel deep self-learning framework to train a robust network on the real noisy datasets without extra supervision. The proposed approach has several appealing benefits. (1) Different from most existing work, it does not rely on any assumption on the distribution of the noisy labels, making it robust to real noises. (2) It does not need extra clean supervision or accessorial network to help training. (3) A self-learning framework is proposed to train the network in an iterative end-to-end manner, which is effective and efficient. Extensive experiments in challenging benchmarks such as Clothing1M and Food101-N show that our approach outperforms its counterparts in all empirical settings.
****

DSConv: Efficient Convolution Operator
Marcelo Gennari do Nascimento, Roger Fawcett, Victor Adrian Prisacariu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5148-5157
Quantization is a popular way of increasing the speed and lowering the memory usage of Convolution Neural Networks (CNNs). When labelled training data is available, network weights and activations have successfully been quantized down to 1-bit. The same cannot be said about the scenario when labelled training data is not available, e.g. when quantizing a pre-trained model, where current approaches show, at best, no loss of accuracy at 8-bit quantizations. We introduce DSConv, a flexible quantized convolution operator that replaces single-precision operations with their far less expensive integer counterparts, while maintaining the probability distributions over both the kernel weights and the outputs. We test our model as a plug-and-play replacement for standard convolution on most popular neural network architectures, ResNet, DenseNet, GoogLeNet, AlexNet and VGG-Net and demonstrate state-of-the-art results, with less than 1% loss of accuracy, without retraining, using only 4-bit quantization. We also show how a distillation-based adaptation stage with unlabelled data can improve results even further.
****

Explicit Shape Encoding for Real-Time Instance Segmentation
Wenqiang Xu, Haiyang Wang, Fubo Qi, Cewu Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5168-5177

In this paper, we propose a novel top-down instance segmentation framework based on explicit shape encoding, named ESE-Seg. It largely reduces the computational consumption of the instance segmentation by explicitly decoding the multiple object shapes with tensor operations, thus performs the instance segmentation at almost the same speed as the object detection. ESE-Seg is based on a novel shape signature Inner-center Radius (IR), Chebyshev polynomial fitting and the strong modern object detectors. ESE-Seg with YOLOv3 outperforms the Mask R-CNN on Pascal VOC 2012 at mAP^r@0.5 while 7 times faster.
****

IMP: Instance Mask Projection for High Accuracy Semantic Segmentation of Things
Cheng-Yang Fu, Tamara L. Berg, Alexander C. Berg; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5178-5187
In this work, we present a new operator, called Instance Mask Projection (IMP), which projects a predicted instance segmentation as a new feature for semantic segmentation. It also supports back propagation and is trainable end-to end. By adding this operator, we introduce a new way to combine top-down and bottom-up information in semantic segmentation. Our experiments show the effectiveness of IMP on both clothing parsing (with complex layering, large deformations, and non-convex objects), and on street scene segmentation (with many overlapping instances and small objects). On the Varied Clothing Parsing dataset (VCP), we show instance mask projection can improve mIOU by 3 points over a state-of-the-art Panoptic FPN segmentation approach. On the ModaNet clothing parsing dataset, we show a dramatic improvement of 20.4% compared to existing baseline semantic segmentation results. In addition, the Instance Mask Projection operator works well on other (non-clothing) datasets, providing an improvement in mIOU of 3 points on "thing" classes of Cityscapes, a self-driving dataset, over a state-of-the-art approach.
****

Video Instance Segmentation
Linjie Yang, Yuchen Fan, Ning Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5188-5197
In this paper we present a new computer vision task, named video instance segmentation. The goal of this new task is simultaneous detection, segmentation and tracking of instances in videos. In words, it is the first time that the image instance segmentation problem is extended to the video domain. To facilitate research on this new task, we propose a large-scale benchmark called YouTube-VIS, which consists of 2,883 high-resolution YouTube videos, a 40-category label set and 131k high-quality instance masks. In addition, we propose a novel algorithm called MaskTrack R-CNN for this task. Our new method introduces a new tracking branch to Mask R-CNN to jointly perform the detection, segmentation and tracking tasks simultaneously. Finally, we evaluate the proposed method and several strong baselines on our new dataset. Experimental results clearly demonstrate the advantages of the proposed algorithm and reveal insight for future improvement. We believe the video instance segmentation task will motivate the community along the line of research for video understanding.
****

Self-Supervised Difference Detection for Weakly-Supervised Semantic Segmentation
Wataru Shimoda, Keiji Yanai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5208-5217
To minimize the annotation costs associated with the training of semantic segmentation models, researchers have extensively investigated weakly-supervised segmentation approaches. In the current weakly-supervised segmentation methods, the most widely adopted approach is based on visualization. However, the visualization results are not generally equal to semantic segmentation. Therefore, to perform accurate semantic segmentation under the weakly supervised condition, it is necessary to consider the mapping functions that convert the visualization results into semantic segmentation. For such mapping functions, the conditional random field and iterative re-training using the outputs of a segmentation model are usually used. However, these methods do not always guarantee improvements in accuracy; therefore, if we apply these mapping functions iteratively multiple times,

eventually the accuracy will not improve or will decrease. In this paper, to mak e the most of such mapping functions, we assume that the results of the mapping function include noise, and we improve the accuracy by removing noise. To achiev e our aim, we propose the self-supervised difference detection module, which est imates noise from the results of the mapping functions by predicting the differe nce between the segmentation masks before and after the mapping. We verified the effectiveness of the proposed method by performing experiments on the PASCAL Vi sual Object Classes 2012 dataset, and we achieved 64.9% in the val set and 65.5% in the test set. Both of the results become new state-of-the-art under the same setting of weakly supervised semantic segmentation.

****

SPGNet: Semantic Prediction Guidance for Scene Parsing

Bowen Cheng, Liang-Chieh Chen, Yunchao Wei, Yukun Zhu, Zilong Huang, Jinjun Xiong, Thomas S. Huang, Wen-Mei Hwu, Honghui Shi; Proceedings of the IEEE/CV F International Conference on Computer Vision (ICCV), 2019, pp. 5218-5228

Multi-scale context module and single-stage encoder-decoder structure are common ly employed for semantic segmentation. The multi-scale context module refers to the operations to aggregate feature responses from a large spatial extent, while the single-stage encoder-decoder structure encodes the high-level semantic info rmation in the encoder path and recovers the boundary information in the decoder path. In contrast, multi-stage encoder-decoder networks have been widely used i n human pose estimation and show superior performance than their single-stage co unterpart. However, few efforts have been attempted to bring this effective desi gn to semantic segmentation. In this work, we propose a Semantic Prediction Guid ance (SPG) module which learns to re-weight the local features through the guida nce from pixel-wise semantic prediction. We find that by carefully re-weighting features across stages, a two-stage encoder-decoder network coupled with our pro posed SPG module can significantly outperform its one-stage counterpart with sim ilar parameters and computations. Finally, we report experimental results on the semantic segmentation benchmark Cityscapes, in which our SPGNet attains 81.1% o n the test set using only 'fine' annotations.

****

Gated-SCNN: Gated Shape CNNs for Semantic Segmentation

Towaki Takikawa, David Acuna, Varun Jampani, Sanja Fidler; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5229-523 8

Current state-of-the-art methods for image segmentation form a dense image repre sentation where the color, shape and texture information are all processed toget her inside a deep CNN. This however may not be ideal as they contain very differ ent type of information relevant for recognition. Here, we propose a new two-str eam CNN architecture for semantic segmentation that explicitly wires shape infor mation as a separate processing branch, i.e. shape stream, that processes inform ation in parallel to the classical stream. Key to this architecture is a new typ e of gates that connect the intermediate layers of the two streams. Specifically , we use the higher-level activations in the classical stream to gate the lower- level activations in the shape stream, effectively removing noise and helping th e shape stream to only focus on processing the relevant boundary-related informa tion. This enables us to use a very shallow architecture for the shape stream th at operates on the image-level resolution. Our experiments show that this leads to a highly effective architecture that produces sharper predictions around obje ct boundaries and significantly boosts performance on thinner and smaller object s. Our method achieves state-of-the-art performance on the Cityscapes benchmark, in terms of both mask (mIoU) and boundary (F-score) quality, improving by 2% an d 4% over strong baselines.

****

DensePoint: Learning Densely Contextual Representation for Efficient Point Cloud Processing

Yongcheng Liu, Bin Fan, Gaofeng Meng, Jiwen Lu, Shiming Xiang, Chunhong Pan ; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) , 2019, pp. 5239-5248

Point cloud processing is very challenging, as the diverse shapes formed by irregular points are often indistinguishable. A thorough grasp of the elusive shape requires sufficiently contextual semantic information, yet few works devote to this. Here we propose DensePoint, a general architecture to learn densely contextual representation for point cloud processing. Technically, it extends regular grid CNN to irregular point configuration by generalizing a convolution operator, which holds the permutation invariance of points, and achieves efficient inductive learning of local patterns. Architecturally, it finds inspiration from dense connection mode, to repeatedly aggregate multi-level and multi-scale semantics in a deep hierarchy. As a result, densely contextual information along with rich semantics, can be acquired by DensePoint in an organic manner, making it highly effective. Extensive experiments on challenging benchmarks across four tasks, as well as thorough model analysis, verify DensePoint achieves the state of the arts.
****

AMP: Adaptive Masked Proxies for Few-Shot Segmentation
Mennatullah Siam, Boris N. Oreshkin, Martin Jagersand; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5249-5258
Deep learning has thrived by training on large-scale datasets. However, in robotics applications sample efficiency is critical. We propose a novel adaptive masked proxies method that constructs the final segmentation layer weights from few labelled samples. It utilizes multi-resolution average pooling on base embeddings masked with the label to act as a positive proxy for the new class, while fusing it with the previously learned class signatures. Our method is evaluated on PASCAL-5^i dataset and outperforms the state-of-the-art in the few-shot semantic segmentation. Unlike previous methods, our approach does not require a second branch to estimate parameters or prototypes, which enables it to be used with 2-stream motion and appearance based segmentation networks. We further propose a novel setup for evaluating continual learning of object segmentation which we name incremental PASCAL (iPASCAL) where our method outperforms the baseline method. Our code is publicly available at https://github.com/MSiam/AdaptiveMaskedProxies.
****

Universal Semi-Supervised Semantic Segmentation
Tarun Kalluri, Girish Varma, Manmohan Chandraker, C.V. Jawahar; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5259-5270
In recent years, the need for semantic segmentation has arisen across several different applications and environments. However, the expense and redundancy of annotation often limits the quantity of labels available for training in any domain, while deployment is easier if a single model works well across domains. In this paper, we pose the novel problem of universal semi-supervised semantic segmentation and propose a solution framework, to meet the dual needs of lower annotation and deployment costs. In contrast to counterpoints such as fine tuning, joint training or unsupervised domain adaptation, universal semi-supervised segmentation ensures that across all domains: (i) a single model is deployed, (ii) unlabeled data is used, (iii) performance is improved, (iv) only a few labels are needed and (v) label spaces may differ. To address this, we minimize supervised as well as within and cross-domain unsupervised losses, introducing a novel feature alignment objective based on pixel-aware entropy regularization for the latter. We demonstrate quantitative advantages over other approaches on several combinations of segmentation datasets across different geographies (Germany, England, India) and environments (outdoors, indoors), as well as qualitative insights on the aligned representations.
****

Accelerate Learning of Deep Hashing With Gradient Attention
Long-Kai Huang, Jianda Chen, Sinno Jialin Pan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5271-5280
Recent years have witnessed the success of learning to hash in fast large-scale image retrieval. As deep learning has shown its superior performance on many computer vision applications, recent designs of learning-based hashing models have

been moving from shallow ones to deep architectures. However, based on our analysis, we find that gradient descent based algorithms used in deep hashing models would potentially cause hash codes of a pair of training instances to be updated towards the directions of each other simultaneously during optimization. In the worst case, the paired hash codes switch their directions after update, and consequently, their corresponding distance in the Hamming space remain unchanged. This makes the overall learning process highly inefficient. To address this issue, we propose a new deep hashing model integrated with a novel gradient attention mechanism. Extensive experimental results on three benchmark datasets show that our proposed algorithm is able to accelerate the learning process and obtain competitive retrieval performance compared with state-of-the-art deep hashing models.

****

SVD: A Large-Scale Short Video Dataset for Near-Duplicate Video Retrieval
Qing-Yuan Jiang, Yi He, Gen Li, Jian Lin, Lei Li, Wu-Jun Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5281-5289

With the explosive growth of video data in real applications, near-duplicate video retrieval (NDVR) has become indispensable and challenging, especially for short videos. However, all existing NDVR datasets are introduced for long videos. Furthermore, most of them are small-scale and lack of diversity due to the high cost of collecting and labeling near-duplicate videos. In this paper, we introduce a large-scale short video dataset, called SVD, for the NDVR task. SVD contains over 500,000 short videos and over 30,000 labeled videos of near-duplicates. We use multiple video mining techniques to construct positive/negative pairs. Furthermore, we design temporal and spatial transformations to mimic user-attack behavior in real applications for constructing more difficult variants of SVD. Experiments show that existing state-of-the-art NDVR methods, including real-value based and hashing based methods, fail to achieve satisfactory performance on this challenging dataset. The release of SVD dataset will foster research and system engineering in the NDVR area. The SVD dataset is available at https://svdbase.github.io.

****

Block Annotation: Better Image Annotation With Sub-Image Decomposition
Hubert Lin, Paul Upchurch, Kavita Bala; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5290-5300

Image datasets with high-quality pixel-level annotations are valuable for semantic segmentation: labelling every pixel in an image ensures that rare classes and small objects are annotated. However, full-image annotations are expensive, with experts spending up to 90 minutes per image. We propose block sub-image annotation as a replacement for full-image annotation. Despite the attention cost of frequent task switching, we find that block annotations can be crowdsourced at higher quality compared to full-image annotation with equal monetary cost using existing annotation tools developed for full-image annotation. Surprisingly, we find that 50% pixels annotated with blocks allows semantic segmentation to achieve equivalent performance to 100% pixels annotated. Furthermore, as little as 12% of pixels annotated allows performance as high as 98% of the performance with dense annotation. In weakly-supervised settings, block annotation outperforms existing methods by 3-4% (absolute) given equivalent annotation time. To recover the necessary global structure for applications such as characterizing spatial context and affordance relationships, we propose an effective method to inpaint block-annotated images with high-quality labels without additional human effort. As such, fewer annotations can also be used for these applications compared to full-image annotation.

****

Probabilistic Deep Ordinal Regression Based on Gaussian Processes
Yanzhu Liu, Fan Wang, Adams Wai Kin Kong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5301-5309

With excellent representation power for complex data, deep neural networks (DNNs) based approaches are state-of-the-art for ordinal regression problem which aim

s to classify instances into ordinal categories. However, DNNs are not able to c
apture uncertainties and produce probabilistic interpretations. As a probabilist
ic model, Gaussian Processes (GPs) on the other hand offers uncertainty informat
ion, which is nonetheless lack of scalability for large datasets. This paper ada
pts traditional GPs regression for ordinal regression problem by using both conj
ugate and non-conjugate ordinal likelihood. Based on that, it proposes a deep ne
ural network with a GPs layer on the top, which is trained end-to-end by the sto
chastic gradient descent method for both neural network parameters and GPs param
eters. The parameters in the ordinal likelihood function are learned as neural n
etwork parameters so that the proposed framework is able to produce fitted likel
ihood functions for training sets and make probabilistic predictions for test po
ints. Experimental results on three real-world benchmarks -- image aesthetics ra
ting, historical image grading and age group estimation -- demonstrate that in t
erms of mean absolute error, the proposed approach outperforms state-of-the-art
ordinal regression approaches and provides the confidence for predictions.
****

Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations

Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, Vicente Ordonez; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5310-5319

In this work, we present a framework to measure and mitigate intrinsic biases wi
th respect to protected variables -such as gender- in visual recognition tasks.
We show that trained models significantly amplify the association of target labe
ls with gender beyond what one would expect from biased datasets. Surprisingly,
we show that even when datasets are balanced such that each label co-occurs equa
lly with each gender, learned models amplify the association between labels and
gender, as much as if data had not been balanced! To mitigate this, we adopt an
adversarial approach to remove unwanted features corresponding to protected vari
ables from intermediate representations in a deep neural network - and provide a
detailed analysis of its effectiveness. Experiments on two datasets: the COCO d
ataset (objects), and the imSitu dataset (actions), show reductions in gender bi
as amplification while maintaining most of the accuracy of the original models.
****

Teacher Guided Architecture Search

Pouya Bashivan, Mark Tensen, James J. DiCarlo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5320-5329

Much of the recent improvement in neural networks for computer vision has result
ed from discovery of new networks architectures. Most prior work has used the pe
rformance of candidate models following limited training to automatically guide
the search in a feasible way. Could further gains in computational efficiency be
achieved by guiding the search via measurements of a high performing network wi
th unknown detailed architecture (e.g. the primate visual system)? As one step t
oward this goal, we use representational similarity analysis to evaluate the sim
ilarity of internal activations of candidate networks with those of a (fixed, hi
gh performing) teacher network. We show that adopting this evaluation metric cou
ld produce up to an order of magnitude in search efficiency over performance-gui
ded methods. Our approach finds a convolutional cell structure with similar perf
ormance as was previously found using other methods but at a total computational
cost that is two orders of magnitude lower than Neural Architecture Search (NAS
) and more than four times lower than progressive neural architecture search (PN
AS). We further show that measurements from only 300 neurons from primate visua
l system provides enough signal to find a network with an Imagenet top-1 error t
hat is significantly lower than that achieved by performance-guided architecture
search alone. These results suggest that representational matching can be used
to accelerate network architecture search in cases where one has access to some
or all of the internal representations of a teacher network of interest, such as
the brain's sensory processing networks.
****

FACSIMILE: Fast and Accurate Scans From an Image in Less Than a Second

David Smith, Matthew Loper, Xiaochen Hu, Paris Mavroidis, Javier Romero; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5330-5339

Current methods for body shape estimation either lack detail or require many images. They are usually architecturally complex and computationally expensive. We propose FACSIMILE (FAX), a method that estimates a detailed body from a single photo, lowering the bar for creating virtual representations of humans. Our approach is easy to implement and fast to execute, making it easily deployable. FAX uses an image-translation network which recovers geometry at the original resolution of the image. Counterintuitively, the main loss which drives FAX is on per-pixel surface normals instead of per-pixel depth, making it possible to estimate detailed body geometry without any depth supervision. We evaluate our approach both qualitatively and quantitatively, and compare with a state-of-the-art method.

****

Delving Deep Into Hybrid Annotations for 3D Human Recovery in the Wild

Yu Rong, Ziwei Liu, Cheng Li, Kaidi Cao, Chen Change Loy; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5340-5348

Though much progress has been achieved in single-image 3D human recovery, estimating 3D model for in-the-wild images remains a formidable challenge. The reason lies in the fact that obtaining high-quality 3D annotations for in-the-wild images is an extremely hard task that consumes enormous amount of resources and manpower. To tackle this problem, previous methods adopt a hybrid training strategy that exploits multiple heterogeneous types of annotations including 3D and 2D while leaving the efficacy of each annotation not thoroughly investigated. In this work, we aim to perform a comprehensive study on cost and effectiveness trade-off between different annotations. Specifically, we focus on the challenging task of in-the-wild 3D human recovery from single images when paired 3D annotations are not fully available. Through extensive experiments, we obtain several observations: 1) 3D annotations are efficient, whereas traditional 2D annotations such as 2D keypoints and body part segmentation are less competent in guiding 3D human recovery. 2) Dense Correspondence such as DensePose is effective. When there are no paired in-the-wild 3D annotations available, the model exploiting dense correspondence can achieve 92% of the performance compared to a model trained with paired 3D data. We show that incorporating dense correspondence into in-the-wild 3D human recovery is promising and competitive due to its high efficiency and relatively low annotating cost. Our model trained with dense correspondence can serve as a strong reference for future research.

****

Human Mesh Recovery From Monocular Images via a Skeleton-Disentangled Representation

Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, Yili Fu, Tao Mei; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5349-5358

We describe an end-to-end method for recovering 3D human body mesh from single images and monocular videos. Different from the existing methods try to obtain all the complex 3D pose, shape, and camera parameters from one coupling feature, we propose a skeleton-disentangling based framework, which divides this task into multi-level spatial and temporal granularity in a decoupling manner. In spatial, we propose an effective and pluggable "disentangling the skeleton from the details" (DSD) module. It reduces the complexity and decouples the skeleton, which lays a good foundation for temporal modeling. In temporal, the self-attention based temporal convolution network is proposed to efficiently exploit the short and long-term temporal cues. Furthermore, an unsupervised adversarial training strategy, temporal shuffles and order recovery, is designed to promote the learning of motion dynamics. The proposed method outperforms the state-of-the-art 3D human mesh recovery methods by 15.4% MPJPE and 23.8% PA-MPJPE on Human3.6M. State-of-the-art results are also achieved on the 3D pose in the wild (3DPW) dataset without any fine-tuning. Especially, ablation studies demonstrate that skeleton-disentangled representation is crucial for better temporal modeling and generaliza

tion.
****
Three-D Safari: Learning to Estimate Zebra Pose, Shape, and Texture From Images "In the Wild"
Silvia Zuffi, Angjoo Kanazawa, Tanya Berger-Wolf, Michael J. Black; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5359-5368
We present the first method to perform automatic 3D pose, shape and texture capture of animals from images acquired in-the-wild. In particular, we focus on the problem of capturing 3D information about Grevy's zebras from a collection of images. The Grevy's zebra is one of the most endangered species in Africa, with only a few thousand individuals left. Capturing the shape and pose of these animals can provide biologists and conservationists with information about animal health and behavior. In contrast to research on human pose, shape and texture estimation, training data for endangered species is limited, the animals are in complex natural scenes with occlusion, they are naturally camouflaged, travel in herds, and look similar to each other. To overcome these challenges, we integrate the recent SMAL animal model into a network-based regression pipeline, which we train end-to-end on synthetically generated images with pose, shape, and background variation. Going beyond state-of-the-art methods for human shape and pose estimation, our method learns a shape space for zebras during training. Learning such a shape space from images using only a photometric loss is novel, and the approach can be used to learn shape in other settings with limited 3D supervision. Moreover, we couple 3D pose and shape prediction with the task of texture synthesis, obtaining a full texture map of the animal from a single image. We show that the predicted texture map allows a novel per-instance unsupervised optimization over the network features. This method, SMALST (SMAL with learned Shape and Texture) goes beyond previous work, which assumed manual keypoints and/or segmentation, to regress directly from pixels to 3D animal shape, pose and texture.
****
Object-Driven Multi-Layer Scene Decomposition From a Single Image
Helisa Dhamo, Nassir Navab, Federico Tombari; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5369-5378
We present a method that tackles the challenge of predicting color and depth behind the visible content of an image. Our approach aims at building up a Layered Depth Image (LDI) from a single RGB input, which is an efficient representation that arranges the scene in layers, including originally occluded regions. Unlike previous work, we enable an adaptive scheme for the number of layers and incorporate semantic encoding for better hallucination of partly occluded objects. Additionally, our approach is object-driven, which especially boosts the accuracy for the occluded intermediate objects. The framework consists of two steps. First, we individually complete each object in terms of color and depth, while estimating the scene layout. Second, we rebuild the scene based on the regressed layers and enforce the recomposed image to resemble the structure of the original input. The learned representation enables various applications, such as 3D photography and diminished reality, all from a single RGB image.
****
Occupancy Flow: 4D Reconstruction by Learning Particle Dynamics
Michael Niemeyer, Lars Mescheder, Michael Oechsle, Andreas Geiger; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5379-5389
Deep learning based 3D reconstruction techniques have recently achieved impressive results. However, while state-of-the-art methods are able to output complex 3D geometry, it is not clear how to extend these results to time-varying topologies. Approaches treating each time step individually lack continuity and exhibit slow inference, while traditional 4D reconstruction methods often utilize a template model or discretize the 4D space at fixed resolution. In this work, we present Occupancy Flow, a novel spatio-temporal representation of time-varying 3D geometry with implicit correspondences. Towards this goal, we learn a temporally and spatially continuous vector field which assigns a motion vector to every poin

t in space and time. In order to perform dense 4D reconstruction from images or sparse point clouds, we combine our method with a continuous 3D representation. Implicitly, our model yields correspondences over time, thus enabling fast infer ence while providing a sound physical description of the temporal dynamics. We s how that our method can be used for interpolation and reconstruction tasks, and demonstrate the accuracy of the learned correspondences. We believe that Occupan cy Flow is a promising new 4D representation which will be useful for a variety of spatio-temporal reconstruction tasks.
****

Joint Monocular 3D Vehicle Detection and Tracking
Hou-Ning Hu, Qi-Zhi Cai, Dequan Wang, Ji Lin, Min Sun, Philipp Krahenbuhl, Trevor Darrell, Fisher Yu; Proceedings of the IEEE/CVF International Conferenc e on Computer Vision (ICCV), 2019, pp. 5390-5399
Vehicle 3D extents and trajectories are critical cues for predicting the future location of vehicles and planning future agent ego-motion based on those predict ions. In this paper, we propose a novel online framework for 3D vehicle detectio n and tracking from monocular videos. The framework can not only associate detec tions of vehicles in motion over time, but also estimate their complete 3D bound ing box information from a sequence of 2D images captured on a moving platform. Our method leverages 3D box depth-ordering matching for robust instance associat ion and utilizes 3D trajectory prediction for re-identification of occluded vehi cles. We also design a motion learning module based on an LSTM for more accurate long-term motion extrapolation. Our experiments on simulation, KITTI, and Argov erse datasets show that our 3D tracking pipeline offers robust data association and tracking. On Argoverse, our image-based method is significantly better for t racking 3D vehicles within 30 meters than the LiDAR-centric baseline methods.
****

Fingerspelling Recognition in the Wild With Iterative Visual Attention
Bowen Shi, Aurora Martinez Del Rio, Jonathan Keane, Diane Brentari, Greg Sha khnarovich, Karen Livescu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5400-5409
Sign language recognition is a challenging gesture sequence recognition problem, characterized by quick and highly coarticulated motion. In this paper we focus on recognition of fingerspelling sequences in American Sign Language (ASL) video s collected in the wild, mainly from YouTube and Deaf social media. Most previou s work on sign language recognition has focused on controlled settings where the data is recorded in a studio environment and the number of signers is limited. Our work aims to address the challenges of real-life data, reducing the need for detection or segmentation modules commonly used in this domain. We propose an e nd-to-end model based on an iterative attention mechanism, without explicit hand detection or segmentation. Our approach dynamically focuses on increasingly hig h-resolution regions of interest. It out-performs prior work by a large margin. We also introduce a newly collected data set of crowdsourced annotations of fing erspelling in the wild, and show that performance can be further improved with t his additional data set.
****

PointAE: Point Auto-Encoder for 3D Statistical Shape and Texture Modelling
Hang Dai, Ling Shao; Proceedings of the IEEE/CVF International Conference on Co mputer Vision (ICCV), 2019, pp. 5410-5419
The outcome of standard statistical shape modelling is a vector space representa tion of objects. Any convex combination of vectors of a set of object class exam ples generates a real and valid example. In this paper, we propose a Point Auto-Encoder (PointAE) with skip-connection, attention blocks for 3D statistical shap e modelling directly on 3D points. The proposed PointAE is able to refine the co rrespondence with a correspondence refinement block. The data with refined corre spondence can be fed to the PointAE again and bootstrap the constructed statisti cal models. Instead of two seperate models, PointAE can simultaneously model the shape and texture variation. The extensive evaluation in three open-sourced dat asets demonstrates that the proposed method achieves better performance in repre sentation ability of the shape variations.

****
Multi-Garment Net: Learning to Dress 3D People From Images
Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, Gerard Pons-Moll; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5420-5430
We present Multi-Garment Network (MGN), a method to predict body shape and clothing, layered on top of the SMPL model from a few frames (1-8) of a video. Several experiments demonstrate that this representation allows higher level of control when compared to single mesh or voxel representations of shape. Our model allows to predict garment geometry, relate it to the body shape, and transfer it to new body shapes and poses. To train MGN, we leverage a digital wardrobe containing 712 digital garments in correspondence, obtained with a novel method to register a set of clothing templates to a dataset of real 3D scans of people in different clothing and poses. Garments from the digital wardrobe, or predicted by MGN, can be used to dress any body shape in arbitrary poses. We will make publicly available the digital wardrobe, the MGN model, and code to dress SMPL with the garments at https://virtualhumans.mpi-inf.mpg.de/mgn
****
Skeleton-Aware 3D Human Shape Reconstruction From Point Clouds
Haiyong Jiang, Jianfei Cai, Jianmin Zheng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5431-5441
This work addresses the problem of 3D human shape reconstruction from point clouds. Considering that human shapes are of high dimensions and with large articulations, we adopt the state-of-the-art parametric human body model, SMPL, to reduce the dimension of learning space and generate smooth and valid reconstruction. However, SMPL parameters, especially pose parameters, are not easy to learn because of ambiguity and locality of the pose representation. Thus, we propose to incorporate skeleton awareness into the deep learning based regression of SMPL parameters for 3D human shape reconstruction. Our basic idea is to use the state-of-the-art technique PointNet++ to extract point features, and then map point features to skeleton joint features and finally to SMPL parameters for the reconstruction from point clouds. Particularly, we develop an end-to-end framework, where we propose a graph aggregation module to augment PointNet++ by extracting better point features, an attention module to better map unordered point features into ordered skeleton joint features, and a skeleton graph module to extract better joint features for SMPL parameter regression. The entire framework network is first trained in an end-to-end manner on synthesized dataset, and then online fine-tuned on unseen dataset with unsupervised loss to bridges gaps between training and testing. The experiments on multiple datasets show that our method is on par with the state-of-the-art solution.
****
AMASS: Archive of Motion Capture As Surface Shapes
Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, Michael J. Black; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5442-5451
Large datasets are the cornerstone of recent advances in computer vision using deep learning. In contrast, existing human motion capture (mocap) datasets are small and the motions limited, hampering progress on learning models of human motion. While there are many different datasets available, they each use a different parameterization of the body, making it difficult to integrate them into a single meta dataset. To address this, we introduce AMASS, a large and varied database of human motion that unifies 15 different optical marker-based mocap datasets by representing them within a common framework and parameterization. We achieve this using a new method, MoSh++, that converts mocap data into realistic 3D human meshes represented by a rigged body model. Here we use SMPL [Loper et al., 2015], which is widely used and provides a standard skeletal representation as well as a fully rigged surface mesh. The method works for arbitrary marker sets, while recovering soft-tissue dynamics and realistic hand motion. We evaluate MoSh++ and tune its hyperparameters using a new dataset of 4D body scans that are jointly recorded with markerbased mocap. The consistent representation of AMASS make

s it readily useful for animation, visualization, and generating training data f
or deep learning. Our dataset is significantly richer than previous human motion
 collections, having more than 40 hours of motion data, spanning over 300 subjec
ts, more than 11000 motions, and will be publicly available to the research comm
unity.
****

Person-in-WiFi: Fine-Grained Person Perception Using WiFi
Fei Wang, Sanping Zhou, Stanislav Panev, Jinsong Han, Dong Huang; Proceeding
s of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp.
5452-5461
Fine-grained person perception such as body segmentation and pose estimation has
 been achieved with many 2D and 3D sensors such as RGB/depth cameras, radars (e.
g. RF-Pose), and LiDARs. These solutions require 2D images, depth maps or 3D poi
nt clouds of person bodies as input. In this paper, we take one step forward to
show that fine-grained person perception is possible even with 1D sensors: WiFi
antennas. Specifically, we used two sets of WiFi antennas to acquire signals, i.
e., one transmitter set and one receiver set. Each set contains three antennas h
orizontally lined-up as a regular household WiFi router. The WiFi signal generat
ed by a transmitter antenna, penetrates through and reflects on human bodies, fu
rniture, and walls, and then superposes at a receiver antenna as 1D signal sampl
es. We developed a deep learning approach that uses annotations on 2D images, ta
kes the received 1D WiFi signals as input, and performs body segmentation and po
se estimation in an end-to-end manner. To our knowledge, our solution is the fir
st work based on off-the-shelf WiFi antennas and standard IEEE 802.11n WiFi sign
als. Demonstrating comparable results to image-based solutions, our WiFi-based p
erson perception solution is cheaper and more ubiquitous than radars and LiDARs,
 while invariant to illumination and has little privacy concern comparing to cam
eras.
****

FAB: A Robust Facial Landmark Detection Framework for Motion-Blurred Videos
Keqiang Sun, Wayne Wu, Tinghao Liu, Shuo Yang, Quan Wang, Qiang Zhou, Zuoc
hang Ye, Chen Qian; Proceedings of the IEEE/CVF International Conference on Com
puter Vision (ICCV), 2019, pp. 5462-5471
Recently, facial landmark detection algorithms have achieved remarkable performa
nce on static images. However, these algorithms are neither accurate nor stable
in motion-blurred videos. The missing of structure information makes it difficul
t for state-of-the-art facial landmark detection algorithms to yield good result
s. In this paper, we propose a framework named FAB that takes advantage of struc
ture consistency in the temporal dimension for facial landmark detection in moti
on-blurred videos. A structure predictor is proposed to predict the missing face
 structural information temporally, which serves as a geometry prior. This allow
s our framework to work as a virtuous circle. On one hand, the geometry prior he
lps our structure-aware deblurring network generates high quality deblurred imag
es which lead to better landmark detection results. On the other hand, better la
ndmark detection results help structure predictor generate better geometry prior
 for the next frame. Moreover, it is a flexible video-based framework that can i
ncorporate any static image-based methods to provide a performance boost on vide
o datasets. Extensive experiments on Blurred-300VW, the proposed Real-world Moti
on Blur (RWMB) datasets and 300VW demonstrate the superior performance to the st
ate-of-the-art methods. Datasets and model will be publicly available at https:
//github.com/KeqiangSun/FAB https://github.com/KeqiangSun/FAB .
****

Attentional Feature-Pair Relation Networks for Accurate Face Recognition
Bong-Nam Kang, Yonghyun Kim, Bongjin Jun, Daijin Kim; Proceedings of the IEEE
/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5472-5481
Human face recognition is one of the most important research areas in biometrics
. However, the robust face recognition under a drastic change of the facial pose
, expression, and illumination is a big challenging problem for its practical ap
plication. Such variations make face recognition more difficult. In this paper,
we propose a novel face recognition method, called Attentional Feature-pair Rela

tion Network (AFRN), which represents the face by the relevant pairs of local ap
pearance block features with their attention scores. The AFRN represents the fac
e by all possible pairs of the 9x9 local appearance block features, the importan
ce of each pair is considered by the attention map that is obtained from the low
-rank bilinear pooling, and each pair is weighted by its corresponding attention
 score. To increase the accuracy, we select top-K pairs of local appearance bloc
k features as relevant facial information and drop the remaining irrelevant. The
 weighted top-K pairs are propagated to extract the joint feature-pair relation
by using bilinear attention network. In experiments, we show the effectiveness o
f the proposed AFRN and achieve the outstanding performance in the 1:1 face veri
fication and 1:N face identification tasks compared to existing state-of-the-art
 methods on the challenging LFW, YTF, CALFW, CPLFW, CFP, AgeDB, IJB-A, IJB-B, an
d IJB-C datasets.
****

Action Recognition With Spatial-Temporal Discriminative Filter Banks
Brais Martinez, Davide Modolo, Yuanjun Xiong, Joseph Tighe; Proceedings of th
e IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5482-54
91
Action recognition has seen a dramatic performance improvement in the last few y
ears. Most of the current state-of-the-art literature either aims at improving p
erformance through changes to the backbone CNN network, or exploring different t
rade-offs between computational efficiency and performance, again through alteri
ng the backbone network. However, almost all of these works maintain the same la
st layers of the network, which simply consist of a global average pooling follo
wed by a fully connected layer. In this work we focus on how to improve the repr
esentation capacity of the network, but rather than altering the backbone, we fo
cus on improving the last layers of the network, where changes have low impact i
n terms of computational cost. In particular, we hypothesize that current archit
ectures have poor sensitivity to finer details and we exploit recent advances in
 the fine-grained recognition literature to improve our model in this aspect. Wi
th the proposed approach, we obtain state-of-the-art performance on Kinetics-400
 and Something-Something-V1, the two major large-scale action recognition benchm
arks.
****

EPIC-Fusion: Audio-Visual Temporal Binding for Egocentric Action Recognition
Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, Dima Damen; Proceedings o
f the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 549
2-5501
We focus on multi-modal fusion for egocentric action recognition, and propose a
novel architecture for multi-modal temporal-binding, i.e. the combination of mod
alities within a range of temporal offsets. We train the architecture with three
 modalities -- RGB, Flow and Audio -- and combine them with mid-level fusion alo
ngside sparse temporal sampling of fused representations. In contrast with previ
ous works, modalities are fused before temporal aggregation, with shared modalit
y fusion weights over time. Our proposed architecture is trained end-to-end, out
performing individual modalities as well as late-fusion of modalities. We demons
trate the importance of audio in egocentric vision, on per-class basis, for iden
tifying actions as well as interacting objects. Our method achieves state of the
 art results on both the seen and unseen test sets of the largest egocentric dat
aset: EPIC-Kitchens, on all metrics using the public leaderboard.
****

Weakly-Supervised Action Localization With Background Modeling
Phuc Xuan Nguyen, Deva Ramanan, Charless C. Fowlkes; Proceedings of the IEEE/C
VF International Conference on Computer Vision (ICCV), 2019, pp. 5502-5511
We describe a latent approach that learns to detect actions in long sequences gi
ven training videos with only whole-video class labels. Our approach makes use o
f two innovations to attention-modeling in weakly-supervised learning. First, an
d most notably, our framework uses an attention model to extract both foreground
 and background frames who's appearance is explicitly modeled. Most prior work i
gnores the background, but we show that modeling it allows our system to learn a

richer notions of actions and their temporal extents. Second, we combine bottom-up, class-agnostic attention modules with top-down, class-specific activation maps, using the latter as form of self-supervision for the former. Doing so allows our model to learn a more accurate model of attention without explicit temporal supervision. These modifications lead to 10% AP@IoU=0.5 improvement over existing systems on THUMOS14. Our proposed weakly-supervised system outperforms the recent state-of-the-art by at least 4.3% AP@IoU=0.5. Finally, we demonstrate that weakly-supervised learning can be used to aggressively scale-up learning to in-the-wild, uncurated Instagram videos (where relevant frames and videos are automatically selected through attentional processing). This allows our weakly supervised approach to even outperform fully-supervised methods for action detection at some overlap thresholds.

****

Grouped Spatial-Temporal Aggregation for Efficient Action Recognition

Chenxu Luo, Alan L. Yuille; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5512-5521

Temporal reasoning is an important aspect of video analysis. 3D CNN shows good performance by exploring spatial-temporal features jointly in an unconstrained way, but it also increases the computational cost a lot. Previous works try to reduce the complexity by decoupling the spatial and temporal filters. In this paper, we propose a novel decomposition method that decomposes the feature channels into spatial and temporal groups in parallel. This decomposition can make two groups focus on static and dynamic cues separately. We call this grouped spatial-temporal aggregation (GST). This decomposition is more parameter-efficient and enables us to quantitatively analyze the contributions of spatial and temporal features in different layers. We verify our model on several action recognition tasks that require temporal reasoning and show its effectiveness.

****

Temporal Structure Mining for Weakly Supervised Action Detection

Tan Yu, Zhou Ren, Yuncheng Li, Enxu Yan, Ning Xu, Junsong Yuan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5522-5531

Different from the fully-supervised action detection problem that is dependent on expensive frame-level annotations, weakly supervised action detection (WSAD) only needs video-level annotations, making it more practical for real-world applications. Existing WSAD methods detect action instances by scoring each video segment (a stack of frames) individually. Most of them fail to model the temporal relations among video segments and cannot effectively characterize action instances possessing latent temporal structure. To alleviate this problem in WSAD, we propose the temporal structure mining (TSM) approach. In TSM, each action instance is modeled as a multi-phase process and phase evolving within an action instance, i.e., the temporal structure, is exploited. Meanwhile, the video background is modeled by a background phase, which separates different action instances in an untrimmed video. In this framework, phase filters are used to calculate the confidence scores of the presence of an action's phases in each segment. Since in the WSAD task, frame-level annotations are not available and thus phase filters cannot be trained directly. To tackle the challenge, we treat each segment's phase as a hidden variable. We use segments' confidence scores from each phase filter to construct a table and determine hidden variables, i.e., phases of segments, by a maximal circulant path discovery along the table. Experiments conducted on three benchmark datasets demonstrate the state-of-the-art performance of the proposed TSM.

****

Temporal Recurrent Networks for Online Action Detection

Mingze Xu, Mingfei Gao, Yi-Ting Chen, Larry S. Davis, David J. Crandall; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5532-5541

Most work on temporal action detection is formulated as an offline problem, in which the start and end times of actions are determined after the entire video is fully observed. However, important real-time applications including surveillanc

e and driver assistance systems require identifying actions as soon as each vide
o frame arrives, based only on current and historical observations. In this pape
r, we propose a novel framework, the Temporal Recurrent Network (TRN), to model
greater temporal context of each frame by simultaneously performing online actio
n detection and anticipation of the immediate future. At each moment in time, ou
r approach makes use of both accumulated historical evidence and predicted futur
e information to better recognize the action that is currently occurring, and in
tegrates both of these into a unified end-to-end architecture. We evaluate our a
pproach on two popular online action detection datasets, HDD and TVSeries, as we
ll as another widely used dataset, THUMOS'14. The results show that TRN signific
antly outperforms the state-of-the-art.
****

StartNet: Online Detection of Action Start in Untrimmed Videos
Mingfei Gao, Mingze Xu, Larry S. Davis, Richard Socher, Caiming Xiong; Proce
edings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019,
pp. 5542-5551
We propose StartNet to address Online Detection of Action Start (ODAS) where act
ion starts and their associated categories are detected in untrimmed, streaming
videos. Previous methods aim to localize action starts by learning feature repre
sentations that can directly separate the start point from its preceding backgro
und. It is challenging due to the subtle appearance difference near the action s
tarts and the lack of training data. Instead, StartNet decomposes ODAS into two
stages: action classification (using ClsNet) and start point localization (using
LocNet). ClsNet focuses on per-frame labeling and predicts action score distrib
utions online. Based on the predicted action scores of the past and current fram
es, LocNet conducts class-agnostic start detection by optimizing long-term local
ization rewards using policy gradient methods. The proposed framework is validat
ed on two large-scale datasets, THUMOS'14 and ActivityNet. The experimental resu
lts show that StartNet significantly outperforms the state-of-the-art by 15%-30%
p-mAP under the offset tolerance of 1-10 seconds on THUMOS'14, and achieves com
parable performance on ActivityNet with 10 times smaller time offset.
****

Video Classification With Channel-Separated Convolutional Networks
Du Tran, Heng Wang, Lorenzo Torresani, Matt Feiszli; Proceedings of the IEEE/
CVF International Conference on Computer Vision (ICCV), 2019, pp. 5552-5561
Group convolution has been shown to offer great computational savings in various
2D convolutional architectures for image classification. It is natural to ask:
1) if group convolution can help to alleviate the high computational cost of vid
eo classification networks; 2) what factors matter the most in 3D group convolut
ional networks; and 3) what are good computation/accuracy trade-offs with 3D gro
up convolutional networks. This paper studies the effects of different design ch
oices in 3D group convolutional networks for video classification. We empiricall
y demonstrate that the amount of channel interactions plays an important role in
the accuracy of 3D group convolutional networks. Our experiments suggest two ma
in findings. First, it is a good practice to factorize 3D convolutions by separa
ting channel interactions and spatiotemporal interactions as this leads to impro
ved accuracy and lower computational cost. Second, 3D channel-separated convolut
ions provide a form of regularization, yielding lower training accuracy but high
er test accuracy compared to 3D convolutions. These two empirical findings lead
us to design an architecture -- Channel-Separated Convolutional Network (CSN) --
which is simple, efficient, yet accurate. On Sports1M and Kinetics, our CSNs ar
e comparable with or better than the state-of-the-art while being 2-3 times more
efficient.
****

Predicting the Future: A Jointly Learnt Model for Action Anticipation
Harshala Gammulle, Simon Denman, Sridha Sridharan, Clinton Fookes; Proceeding
s of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp.
5562-5571
Inspired by human neurological structures for action anticipation, we present an
action anticipation model that enables the prediction of plausible future actio

ns by forecasting both the visual and temporal future. In contrast to current st ate-of-the-art methods which first learn a model to predict future video feature s and then perform action anticipation using these features, the proposed framew ork jointly learns to perform the two tasks, future visual and temporal represen tation synthesis, and early action anticipation. The joint learning framework en sures that the predicted future embeddings are informative to the action anticip ation task. Furthermore, through extensive experimental evaluations we demonstra te the utility of using both visual and temporal semantics of the scene, and ill ustrate how this representation synthesis could be achieved through a recurrent Generative Adversarial Network (GAN) framework. Our model outperforms the curren t state-of-the-art methods on multiple datasets: UCF101, UCF101-24, UT-Interacti on and TV Human Interaction.

****

Human-Aware Motion Deblurring

Ziyi Shen, Wenguan Wang, Xiankai Lu, Jianbing Shen, Haibin Ling, Tingfa Xu, Ling Shao; Proceedings of the IEEE/CVF International Conference on Computer Vi sion (ICCV), 2019, pp. 5572-5581

This paper proposes a human-aware deblurring model that disentangles the motion blur between foreground (FG) humans and background (BG). The proposed model is b ased on a triple-branch encoder-decoder architecture. The first two branches are learned for sharpening FG humans and BG details, respectively; while the third one produces global, harmonious results by comprehensively fusing multi-scale de blurring information from the two domains. The proposed model is further endowed with a supervised, human-aware attention mechanism in an end-to-end fashion. It learns a soft mask that encodes FG human information and explicitly drives the FG/BG decoder-branches to focus on their specific domains. Above designs lead to a fully differentiable motion deblurring network, which can be trained end-to-e nd. To further benefit the research towards Human-aware Image Deblurring, we int roduce a large-scale dataset, named HIDE, which consists of 8,422 blurry and sha rp image pairs with 65,784 densely annotated FG human bounding boxes. HIDE is sp ecifically built to span a broad range of scenes, human object sizes, motion pat terns, and background complexities. Extensive experiments on public benchmarks a nd our dataset demonstrate that our model performs favorably against the state-o f-the-art motion deblurring methods, especially in capturing semantic details.

****

Fast Video Object Segmentation via Dynamic Targeting Network

Lu Zhang, Zhe Lin, Jianming Zhang, Huchuan Lu, You He; Proceedings of the IE EE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5582-5591

We propose a new model for fast and accurate video object segmentation. It consi sts of two convolutional neural networks, a Dynamic Targeting Network (DTN) and a Mask Refinement Network (MRN). DTN locates the object by dynamically focusing on regions of interest surrounding the target object. The target region is predi cted by DTN via two sub-streams, Box Propagation (BP) and Box Re-identification (BR). The BP stream is faster but less effective at objects with large deformati on or occlusion. The BR stream performs better in difficult scenarios at a highe r computation cost. We propose a Decision Module (DM) to adaptively determine wh ich sub-stream to use for each frame. Finally, MRN is exploited to predict segme ntation within the target region. Experimental results on two public datasets de monstrate that the proposed model significantly outperforms existing methods wit hout online training in both accuracy and efficiency, and is comparable to onlin e training-based methods in accuracy with an order of magnitude faster speed.

****

Solving Vision Problems via Filtering

Sean I. Young, Aous T. Naman, Bernd Girod, David Taubman; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5592-5601

We propose a new, filtering approach for solving a large number of regularized i nverse problems commonly found in computer vision. Traditionally, such problems are solved by finding the solution to the system of equations that expresses the first-order optimality conditions of the problem. This can be slow if the syste m of equations is dense due to the use of nonlocal regularization, necessitating

iterative solvers such as successive over-relaxation or conjugate gradients. In this paper, we show that similar solutions can be obtained more easily via filtering, obviating the need to solve a potentially dense system of equations using slow iterative methods. Our filtered solutions are very similar to the true ones, but often up to 10 times faster to compute.

****

GAN-Based Projector for Faster Recovery With Convergence Guarantees in Linear Inverse Problems

Ankit Raj, Yuqi Li, Yoram Bresler; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5602-5611

A Generative Adversarial Network (GAN) with generator G trained to model the prior of images has been shown to perform better than sparsity-based regularizers in ill-posed inverse problems. Here, we propose a new method of deploying a GAN-based prior to solve linear inverse problems using projected gradient descent (PGD). Our method learns a network-based projector for use in the PGD algorithm, eliminating expensive computation of the Jacobian of G. Experiments show that our approach provides a speed-up of 60-80x over earlier GAN-based recovery methods along with better accuracy in compressed sensing. Our main theoretical result is that if the measurement matrix is moderately conditioned on the manifold range(G) and the projector is d-approximate, then the algorithm is guaranteed to reach O(d) reconstruction error in O(log(1/d)) steps in the low noise regime. Additionally, we propose a fast method to design such measurement matrices for a given G. Extensive experiments demonstrate the efficacy of this method by requiring 5-10x fewer measurements than random Gaussian measurement matrices for comparable recovery performance. Because the learning of the GAN and projector is decoupled from the measurement operator, our GAN-based projector and recovery algorithm are applicable without retraining to all linear inverse problems in which the measurement operator is moderately conditioned for range(G), as confirmed by experiments on compressed sensing, super-resolution, and inpainting.

****

Scoot: A Perceptual Metric for Facial Sketches

Deng-Ping Fan, ShengChuan Zhang, Yu-Huan Wu, Yun Liu, Ming-Ming Cheng, Bo Ren, Paul L. Rosin, Rongrong Ji; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5612-5622

While it is trivial for humans to quickly assess the perceptual similarity between two images, the underlying mechanism are thought to be quite complex. Despite this, the most widely adopted perceptual metrics today, such as SSIM and FSIM, are simple, shallow functions, and fail to consider many factors of human perception. Recently, the facial modeling community has observed that the inclusion of both structure and texture has a significant positive benefit for face sketch synthesis (FSS). But how perceptual are these so-called "perceptual features"? Which elements are critical for their success? In this paper, we design a perceptual metric, called Structure Co-Occurrence Texture (Scoot), which simultaneously considers the block-level spatial structure and co-occurrence texture statistics. To test the quality of metrics, we propose three novel meta-measures based on various reliable properties. Extensive experiments verify that our Scoot metric exceeds the performance of prior work. Besides, we built the first largest scale (152k judgments) human-perception-based sketch database that can evaluate how well a metric consistent with human perception. Our results suggest that "spatial structure" and "co-occurrence texture" are two generally applicable perceptual features in face sketch synthesis.

****

Learning Filter Basis for Convolutional Neural Network Compression

Yawei Li, Shuhang Gu, Luc Van Gool, Radu Timofte; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5623-5632

Convolutional neural networks (CNNs) based solutions have achieved state-of-the-art performances for many computer vision tasks, including classification and super-resolution of images. Usually the success of these methods comes with a cost of millions of parameters due to stacking deep convolutional layers. Moreover, quite a large number of filters are also used for a single convolutional layer,

which exaggerates the parameter burden of current methods. Thus, in this paper, we try to reduce the number of parameters of CNNs by learning a basis of the filters in convolutional layers. For the forward pass, the learned basis is used to approximate the original filters and then used as parameters for the convolutional layers. We validate our proposed solution for multiple CNN architectures on image classification and image super-resolution benchmarks and compare favorably to the existing state-of-the-art in terms of reduction of parameters and preservation of accuracy. Code is available at https://github.com/ofsoundof/learning_filter_basis

****

End-to-End Learning of Representations for Asynchronous Event-Based Data
Daniel Gehrig, Antonio Loquercio, Konstantinos G. Derpanis, Davide Scaramuzza; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5633-5643
Event cameras are vision sensors that record asynchronous streams of per-pixel brightness changes, referred to as "events". They have appealing advantages over frame based cameras for computer vision, including high temporal resolution, high dynamic range, and no motion blur. Due to the sparse, non-uniform spatio-temporal layout of the event signal, pattern recognition algorithms typically aggregate events into a grid-based representation and subsequently process it by a standard vision pipeline, e.g., Convolutional Neural Network (CNN). In this work, we introduce a general framework to convert event streams into grid-based representations by means of strictly differentiable operations. Our framework comes with two main advantages: (i) allows learning the input event representation together with the task dedicated network in an end to end manner, and (ii) lays out a taxonomy that unifies the majority of extant event representations in the literature and identifies novel ones. Empirically, we show that our approach to learning the event representation end-to-end yields an improvement of approximately 12% on optical flow estimation and object recognition over state-of-the-art methods.

****

ERL-Net: Entangled Representation Learning for Single Image De-Raining
Guoqing Wang, Changming Sun, Arcot Sowmya; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5644-5652
Despite the significant progress achieved in image de-raining by training an encoder-decoder network within the image-to-image translation formulation, blurry results with missing details indicate the deficiency of the existing models. By interpreting the de-raining encoder-decoder network as a conditional generator, within which the decoder acts as a generator conditioned on the embedding learned by the encoder, the unsatisfactory output can be attributed to the low-quality embedding learned by the encoder. In this paper, we hypothesize that there exists an inherent mapping between the low-quality embedding to a latent optimal one, with which the generator (decoder) can produce much better results. To improve the de-raining results significantly over existing models, we propose to learn this mapping by formulating a residual learning branch, that is capable of adaptively adding residuals to the original low-quality embedding in a representation entanglement manner. Using an embedding learned this way, the decoder is able to generate much more satisfactory de-raining results with better detail recovery and rain artefacts removal, providing new state-of-the-art results on four benchmark datasets with considerable improvement (i.e., on the challenging Rain100H data, an improvement of 4.19dB on PSNR and 5% on SSIM is obtained). The entanglement can be easily adopted into any encoder-decoder based image restoration networks. Besides, we propose a series of evaluation metrics to investigate the specific contribution of the proposed entangled representation learning mechanism. Codes are available at .

****

Perceptual Deep Depth Super-Resolution
Oleg Voynov, Alexey Artemov, Vage Egiazarian, Alexander Notchenko, Gleb Bobrovskikh, Evgeny Burnaev, Denis Zorin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5653-5663

RGBD images, combining high-resolution color and lower-resolution depth from various types of depth sensors, are increasingly common. One can significantly improve the resolution of depth maps by taking advantage of color information; deep learning methods make combining color and depth information particularly easy. However, fusing these two sources of data may lead to a variety of artifacts. If depth maps are used to reconstruct 3D shapes, e.g., for virtual reality applications, the visual quality of upsampled images is particularly important. The main idea of our approach is to measure the quality of depth map upsampling using renderings of resulting 3D surfaces. We demonstrate that a simple visual appearance-based loss, when used with either a trained CNN or simply a deep prior, yields significantly improved 3D shapes, as measured by a number of existing perceptual metrics. We compare this approach with a number of existing optimization and learning-based techniques.

****

3D Scene Graph: A Structure for Unified Semantics, 3D Space, and Camera

Iro Armeni, Zhi-Yang He, JunYoung Gwak, Amir R. Zamir, Martin Fischer, Jitendra Malik, Silvio Savarese; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5664-5673

A comprehensive semantic understanding of a scene is important for many applications - but in what space should diverse semantic information (e.g., objects, scene categories, material types, 3D shapes, etc.) be grounded and what should be its structure? Aspiring to have one unified structure that hosts diverse types of semantics, we follow the Scene Graph paradigm in 3D, generating a 3D Scene Graph. Given a 3D mesh and registered panoramic images, we construct a graph that spans the entire building and includes semantics on objects (e.g., class, material, shape and other attributes), rooms (e.g., function, illumination type, etc.) and cameras (e.g., location, etc.), as well as the relationships among these entities. However, this process is prohibitively labor heavy if done manually. To alleviate this we devise a semi-automatic framework that employs existing detection methods and enhances them using two main constraints: I. framing of query images sampled on panoramas to maximize the performance of 2D detectors, and II. multi-view consistency enforcement across 2D detections that originate in different camera locations.

****

Floorplan-Jigsaw: Jointly Estimating Scene Layout and Aligning Partial Scans

Cheng Lin, Changjian Li, Wenping Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5674-5683

We present a novel approach to align partial 3D reconstructions which may not have substantial overlap. Using floorplan priors, our method jointly predicts a room layout and estimates the transformations from a set of partial 3D data. Unlike the existing methods relying on feature descriptors to establish correspondences, we exploit the 3D "box" structure of a typical room layout that meets the Manhattan World property. We first estimate a local layout for each partial scan separately and then combine these local layouts to form a globally aligned layout with loop closure. Without the requirement of feature matching, the proposed method enables some novel applications ranging from large or featureless scene reconstruction and modeling from sparse input. We validate our method quantitatively and qualitatively on real and synthetic scenes of various sizes and complexities. The evaluations and comparisons show superior effectiveness and accuracy of our method.

****

Enforcing Geometric Constraints of Virtual Normal for Depth Prediction

Wei Yin, Yifan Liu, Chunhua Shen, Youliang Yan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5684-5693

Monocular depth prediction plays a crucial role in understanding 3D scene geometry. Although recent methods have achieved impressive progress in evaluation metrics such as the pixel-wise relative error, most methods neglect the geometric constraints in the 3D space. In this work, we show the importance of the high-order 3D geometric constraints for depth prediction. By designing a loss term that enforces one simple type of geometric constraints, namely, virtual normal directi

ons determined by randomly sampled three points in the reconstructed 3D space, we can considerably improve the depth prediction accuracy. Furthermore, we can not only predict accurate depth but also achieve high-quality other 3D information from the depth without retraining new parameters, Significantly, the byproduct of this predicted depth being sufficiently accurate is that we are now able to recover good 3D structures of the scene such as the point cloud and surface normal directly from the depth, eliminating the necessity of training new sub-models as was previously done. Experiments on two challenging benchmarks: NYU Depth-V2 and KITTI demonstrate the effectiveness of our method and state-of-the-art performance.

****

Deep Contextual Attention for Human-Object Interaction Detection

Tiancai Wang, Rao Muhammad Anwer, Muhammad Haris Khan, Fahad Shahbaz Khan, Yanwei Pang, Ling Shao, Jorma Laaksonen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5694-5702

Human-object interaction detection is an important and relatively new class of visual relationship detection tasks, essential for deeper scene understanding. Most existing approaches decompose the problem into object localization and interaction recognition. Despite showing progress, these approaches only rely on the appearances of humans and objects and overlook the available context information, crucial for capturing subtle interactions between them. We propose a contextual attention framework for human-object interaction detection. Our approach leverages context by learning contextually-aware appearance features for human and object instances. The proposed attention module then adaptively selects relevant instance-centric context information to highlight image regions likely to contain human-object interactions. Experiments are performed on three benchmarks: V-COCO, HICO-DET and HCVRD. Our approach outperforms the state-of-the-art on all datasets. On the V-COCO dataset, our method achieves a relative gain of 4.4% in terms of role mean average precision (mAP role), compared to the existing best approach.

****

Learning Compositional Neural Information Fusion for Human Parsing

Wenguan Wang, Zhijie Zhang, Siyuan Qi, Jianbing Shen, Yanwei Pang, Ling Shao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5703-5713

This work proposes to combine neural networks with the compositional hierarchy of human bodies for efficient and complete human parsing. We formulate the approach as a neural information fusion framework. Our model assembles the information from three inference processes over the hierarchy: direct inference (directly predicting each part of a human body using image information), bottom-up inference (assembling knowledge from constituent parts), and top-down inference (leveraging context from parent nodes). The bottom-up and top-down inferences explicitly model the compositional and decompositional relations in human bodies, respectively. In addition, the fusion of multi-source information is conditioned on the inputs, i.e., by estimating and considering the confidence of the sources. The whole model is end-to-end differentiable, explicitly modeling information flows and structures. Our approach is extensively evaluated on four popular datasets, outperforming the state-of-the-arts in all cases, with a fast processing speed of 23fps. Our code and results have been released to help ease future research in this direction.

****

Attentional Neural Fields for Crowd Counting

Anran Zhang, Lei Yue, Jiayi Shen, Fan Zhu, Xiantong Zhen, Xianbin Cao, Ling Shao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5714-5723

Crowd counting has recently generated huge popularity in computer vision, and is extremely challenging due to the huge scale variations of objects. In this paper, we propose the Attentional Neural Field (ANF) for crowd counting via density estimation. Within the encoder-decoder network, we introduce conditional random fields (CRFs) to aggregate multi-scale features, which can build more informativ

e representations. To better model pair-wise potentials in CRFs, we incorperate non-local attention mechanism implemented as inter- and intra-layer attentions to expand the receptive field to the entire image respectively within the same layer and across different layers, which captures long-range dependencies to conquer huge scale variations. The CRFs coupled with the attention mechanism are seamlessly integrated into the encoder-decoder network, establishing an ANF that can be optimized end-to-end by back propagation. We conduct extensive experiments on four public datasets, including ShanghaiTech, WorldEXPO 10, UCF-CC-50 and UCF-QNRF. The results show that our ANF achieves high counting performance, surpassing most previous methods.

****

Understanding Human Gaze Communication by Spatio-Temporal Graph Reasoning

Lifeng Fan, Wenguan Wang, Siyuan Huang, Xinyu Tang, Song-Chun Zhu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5724-5733

This paper addresses a new problem of understanding human gaze communication in social videos from both atomic-level and event-level, which is significant for studying human social interactions. To tackle this novel and challenging problem, we contribute a large-scale video dataset, VACATION, which covers diverse daily social scenes and gaze communication behaviors with complete annotations of objects and human faces, human attention, and communication structures and labels in both atomic-level and event-level. Together with VACATION, we propose a spatio-temporal graph neural network to explicitly represent the diverse gaze interactions in the social scenes and to infer atomic-level gaze communication by message passing. We further propose an event network with encoder-decoder structure to predict the event-level gaze communication. Our experiments demonstrate that the proposed model improves various baselines significantly in predicting the atomic-level and event-level gaze communications.

****

Controllable Attention for Structured Layered Video Decomposition

Jean-Baptiste Alayrac, Joao Carreira, Relja Arandjelovic, Andrew Zisserman; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5734-5743

The objective of this paper is to be able to separate a video into its natural layers, and to control which of the separated layers to attend to. For example, to be able to separate reflections, transparency or object motion. We make the following three contributions: (i) we introduce a new structured neural network architecture that explicitly incorporates layers (as spatial masks) into its design. This improves separation performance over previous general purpose networks for this task; (ii) we demonstrate that we can augment the architecture to leverage external cues such as audio for controllability and to help disambiguation; and (iii) we experimentally demonstrate the effectiveness of our approach and training procedure with controlled experiments while also showing that the proposed model can be successfully applied to real-word applications such as reflection removal and action recognition in cluttered scenes.

****

GANalyze: Toward Visual Definitions of Cognitive Image Properties

Lore Goetschalckx, Alex Andonian, Aude Oliva, Phillip Isola; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5744-5753

We introduce a framework that uses Generative Adversarial Networks (GANs) to study cognitive properties like memorability. These attributes are of interest because we do not have a concrete visual definition of what they entail. What does it look like for a dog to be more memorable? GANs allow us to generate a manifold of natural-looking images with fine-grained differences in their visual attributes. By navigating this manifold in directions that increase memorability, we can visualize what it looks like for a particular generated image to become more memorable. The resulting "visual definitions" surface image properties (like "object size") that may underlie memorability. Through behavioral experiments, we verify that our method indeed discovers image manipulations that causally affect h

uman memory performance. We further demonstrate that the same framework can be u
sed to analyze image aesthetics and emotional valence. ganalyze.csail.mit.edu.
****

Saliency-Guided Attention Network for Image-Sentence Matching
Zhong Ji, Haoran Wang, Jungong Han, Yanwei Pang; Proceedings of the IEEE/CVF
International Conference on Computer Vision (ICCV), 2019, pp. 5754-5763
This paper studies the task of matching image and sentence, where learning appro
priate representations to bridge the semantic gap between image contents and lan
guage appears to be the main challenge. Unlike previous approaches that predomin
antly deploy symmetrical architecture to represent both modalities, we introduce
a Saliency-guided Attention Network (SAN) that is characterized by building an
asymmetrical link between vision and language to efficiently learn a fine-graine
d cross-modal correlation. The proposed SAN mainly includes three components: sa
liency detector, Saliency-weighted Visual Attention (SVA) module, and Saliency-g
uided Textual Attention (STA) module. Concretely, the saliency detector provides
the visual saliency information to drive both two attention modules. Taking adv
antage of the saliency information, SVA is able to learn more discriminative vis
ual features. By fusing the visual information from SVA and intra-modal informat
ion as a multi-modal guidance, STA affords us powerful textual representations t
hat are synchronized with visual clues. Extensive experiments demonstrate SAN ca
n improve the state-of-the-art results on the benchmark Flickr30K and MSCOCO dat
asets by a large margin.
****

CAMP: Cross-Modal Adaptive Message Passing for Text-Image Retrieval
Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang,
Jing Shao; Proceedings of the IEEE/CVF International Conference on Computer Visi
on (ICCV), 2019, pp. 5764-5773
Text-image cross-modal retrieval is a challenging task in the field of language
and vision. Most previous approaches independently embed images and sentences in
to a joint embedding space and compare their similarities. However, previous app
roaches rarely explore the interactions between images and sentences before calc
ulating similarities in the joint space. Intuitively, when matching between imag
es and sentences, human beings would alternatively attend to regions in images a
nd words in sentences, and select the most salient information considering the i
nteraction between both modalities. In this paper, we propose Cross-modal Adapti
ve Message Passing (CAMP), which adaptively controls the information flow for me
ssage passing across modalities. Our approach not only takes comprehensive and f
ine-grained cross-modal interactions into account, but also properly handles neg
ative pairs and irrelevant information with an adaptive gating scheme. Moreover,
instead of conventional joint embedding approaches for text-image matching, we
infer the matching score based on the fused features, and propose a hardest nega
tive binary cross-entropy loss for training. Results on COCO and Flickr30k signi
ficantly surpass state-of-the-art methods, demonstrating the effectiveness of ou
r approach.
****

ACMM: Aligned Cross-Modal Memory for Few-Shot Image and Sentence Matching
Yan Huang, Liang Wang; Proceedings of the IEEE/CVF International Conference on
Computer Vision (ICCV), 2019, pp. 5774-5783
Image and sentence matching has drawn much attention recently, but due to the la
ck of sufficient pairwise data for training, most previous methods still cannot
well associate those challenging pairs of images and sentences containing rarely
appeared regions and words, i.e., few-shot content. In this work, we study this
challenging scenario as few-shot image and sentence matching, and accordingly p
ropose an Aligned Cross-Modal Memory (ACMM) model to memorize the rarely appeare
d content. Given a pair of image and sentence, the model first includes an align
ed memory controller network to produce two sets of semantically-comparable inte
rface vectors through cross-modal alignment. Then the interface vectors are used
by modality-specific read and update operations to alternatively interact with
shared memory items. The memory items persistently memorize cross-modal shared s
emantic representations, which can be addressed out to better enhance the repres

entation of few-shot content. We apply the proposed model to both conventional and few-shot image and sentence matching tasks, and demonstrate its effectiveness by achieving the state-of-the-art performance on two benchmark datasets.
****

Creativity Inspired Zero-Shot Learning
Mohamed Elhoseiny, Mohamed Elfeki; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5784-5793
Zero-shot learning (ZSL) aims at understanding unseen categories with no training examples from class-level descriptions. To improve the discriminative power of zero-shot learning, we model the visual learning process of unseen categories with an inspiration from the psychology of human creativity for producing novel art. We relate ZSL to human creativity by observing that zero-shot learning is about recognizing the unseen and creativity is about creating a likable unseen. We introduce a learning signal inspired by creativity literature that explores the unseen space with hallucinated class-descriptions and encourages careful deviation of their visual feature generations from seen classes while allowing knowledge transfer from seen to unseen classes. Empirically, we show consistent improvement over the state of the art of several percents on the largest available benchmarks on the challenging task or generalized ZSL from a noisy text that we focus on, using the CUB and NABirds datasets. We also show the advantage of our loss on Attribute-based ZSL on three additional datasets (AwA2, aPY, and SUN). Code is available at https://github.com/mhelhoseiny/CIZSL.
****

Generating Easy-to-Understand Referring Expressions for Target Identifications
Mikihiro Tanaka, Takayuki Itamochi, Kenichi Narioka, Ikuro Sato, Yoshitaka Ushiku, Tatsuya Harada; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5794-5803
This paper addresses the generation of referring expressions that not only refer to objects correctly but also let humans find them quickly. As a target becomes relatively less salient, identifying referred objects itself becomes more difficult. However, the existing studies regarded all sentences that refer to objects correctly as equally good, ignoring whether they are easily understood by humans. If the target is not salient, humans utilize relationships with the salient contexts around it to help listeners to comprehend it better. To derive this information from human annotations, our model is designed to extract information from the target and from the environment. Moreover, we regard that sentences that are easily understood are those that are comprehended correctly and quickly by humans. We optimized this by using the time required to locate the referred objects by humans and their accuracies. To evaluate our system, we created a new referring expression dataset whose images were acquired from Grand Theft Auto V (GTA V), limiting targets to persons. Experimental results show the effectiveness of our approach. Our code and dataset are available at https://github.com/mikittt/easy-to-understand-REG.
****

Language-Agnostic Visual-Semantic Embeddings
Jonatas Wehrmann, Douglas M. Souza, Mauricio A. Lopes, Rodrigo C. Barros; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5804-5813
This paper proposes a framework for training language-invariant cross-modal retrieval models. We also introduce a novel character-based word-embedding approach, allowing the model to project similar words across languages into the same word-embedding space. In addition, by performing cross-modal retrieval at the character level, the storage requirements for a text encoder decrease substantially, allowing for lighter and more scalable retrieval architectures. The proposed language-invariant textual encoder based on characters is virtually unaffected in terms of storage requirements when novel languages are added to the system. Our contributions include new methods for building character-level-based word-embeddings, an improved loss function, and a novel cross-language alignment module that not only makes the architecture language-invariant, but also presents better predictive performance. We show that our models outperform the current state-of-the

-art in both single and multi-language scenarios. This work can be seen as the b
asis of a new path on retrieval research, now allowing for the effective use of
captions in multiple-language scenarios. Code is available at https://github.com
/jwehrmann/lavse.
****

Adversarial Representation Learning for Text-to-Image Matching
Nikolaos Sarafianos, Xiang Xu, Ioannis A. Kakadiaris; Proceedings of the IEEE/
CVF International Conference on Computer Vision (ICCV), 2019, pp. 5814-5824
For many computer vision applications such as image captioning, visual question
answering, and person search, learning discriminative feature representations at
both image and text level is an essential yet challenging problem. Its challeng
es originate from the large word variance in the text domain as well as the diff
iculty of accurately measuring the distance between the features of the two moda
lities. Most prior work focuses on the latter challenge, by introducing loss fun
ctions that help the network learn better feature representations but fail to ac
count for the complexity of the textual input. With that in mind, we introduce T
IMAM: a Text-Image Modality Adversarial Matching approach that learns modality-i
nvariant feature representations using adversarial and cross-modal matching obje
ctives. In addition, we demonstrate that BERT, a publicly-available language mod
el that extracts word embeddings, can successfully be applied in the text-to-ima
ge matching domain. The proposed approach achieves state-of-the-art cross-modal
matching performance on four widely-used publicly-available datasets resulting i
n absolute improvements ranging from 2% to 5% in terms of rank-1 accuracy.
****

Multi-Modality Latent Interaction Network for Visual Question Answering
Peng Gao, Haoxuan You, Zhanpeng Zhang, Xiaogang Wang, Hongsheng Li; Proceedi
ngs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp
. 5825-5835
Exploiting relationships between visual regions and question words have achieved
great success in learning multi-modality features for Visual Question Answering
(VQA). However, we argue that existing methods mostly model relations between i
ndividual visual regions and words, which are not enough to correctly answer the
question. From humans' perspective, answering a visual question requires unders
tanding the summarizations of visual and language information. In this paper, we
proposed the Multi-modality Latent Interaction module (MLI) to tackle this prob
lem. The proposed module learns the cross-modality relationships between latent
visual and language summarizations, which summarize visual regions and question
into a small number of latent representations to avoid modeling uninformative in
dividual region-word relations. The cross-modality information between the laten
t summarizations are propagated to fuse valuable information from both modalitie
s and are used to update the visual and word features. Such MLI modules can be s
tacked for several stages to model complex and latent relations between the two
modalities and achieves highly competitive performance on public VQA benchmarks,
VQA v2.0 and TDIUC . In addition, we show that the performance of our methods c
ould be significantly improved by combining with pre-trained language model BERT
.
****

Key.Net: Keypoint Detection by Handcrafted and Learned CNN Filters
Axel Barroso-Laguna, Edgar Riba, Daniel Ponsa, Krystian Mikolajczyk; Proceedi
ngs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp
. 5836-5844
We introduce a novel approach for keypoint detection task that combines handcraf
ted and learned CNN filters within a shallow multi-scale architecture. Handcraft
ed filters provide anchor structures for learned filters, which localize, score
and rank repeatable features. Scale-space representation is used within the netw
ork to extract keypoints at different levels. We design a loss function to detec
t robust features that exist across a range of scales and to maximize the repeat
ability score. Our Key.Net model is trained on data synthetically created from I
mageNet and evaluated on HPatches benchmark. Results show that our approach outp
erforms state-of-the-art detectors in terms of repeatability, matching performan

ce and complexity.
****

Learning Two-View Correspondences and Geometry Using Order-Aware Network
Jiahui Zhang, Dawei Sun, Zixin Luo, Anbang Yao, Lei Zhou, Tianwei Shen, Yurong Chen, Long Quan, Hongen Liao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5845-5854
Establishing correspondences between two images requires both local and global spatial context. Given putative correspondences of feature points in two views, in this paper, we propose Order-Aware Network, which infers the probabilities of correspondences being inliers and regresses the relative pose encoded by the essential matrix. Specifically, this proposed network is built hierarchically and comprises three novel operations. First, to capture the local context of sparse correspondences, the network clusters unordered input correspondences by learning a soft assignment matrix. These clusters are in a canonical order and invariant to input permutations. Next, the clusters are spatially correlated to form the global context of correspondences. After that, the context-encoded clusters are recovered back to the original size through a proposed upsampling operator. We intensively experiment on both outdoor and indoor datasets. The accuracy of the two-view geometry and correspondences are significantly improved over the state-of-the-arts.
****

Learning Meshes for Dense Visual SLAM
Michael Bloesch, Tristan Laidlow, Ronald Clark, Stefan Leutenegger, Andrew J. Davison; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5855-5864
Estimating motion and surrounding geometry of a moving camera remains a challenging inference problem. From an information theoretic point of view, estimates should get better as more information is included, such as is done in dense SLAM, but this is strongly dependent on the validity of the underlying models. In the present paper, we use triangular meshes as both compact and dense geometry representation. To allow for simple and fast usage, we propose a view-based formulation for which we predict the in-plane vertex coordinates directly from images and then employ the remaining vertex depth components as free variables. Flexible and continuous integration of information is achieved through the use of a residual based inference technique. This so-called factor graph encodes all information as mapping from free variables to residuals, the squared sum of which is minimised during inference. We propose the use of different types of learnable residuals, which are trained end-to-end to increase their suitability as information bearing models and to enable accurate and reliable estimation. Detailed evaluation of all components is provided on both synthetic and real data which confirms the practicability of the presented approach.
****

EM-Fusion: Dynamic Object-Level SLAM With Probabilistic Data Association
Michael Strecke, Jorg Stuckler; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5865-5874
The majority of approaches for acquiring dense 3D environment maps with RGB-D cameras assumes static environments or rejects moving objects as outliers. The representation and tracking of moving objects, however, has significant potential for applications in robotics or augmented reality. In this paper, we propose a novel approach to dynamic SLAM with dense object-level representations. We represent rigid objects in local volumetric signed distance function (SDF) maps, and formulate multi-object tracking as direct alignment of RGB-D images with the SDF representations. Our main novelty is a probabilistic formulation which naturally leads to strategies for data association and occlusion handling. We analyze our approach in experiments and demonstrate that our approach compares favorably with the state-of-the-art methods in terms of robustness and accuracy.
****

ClusterSLAM: A SLAM Backend for Simultaneous Rigid Body Clustering and Motion Estimation
Jiahui Huang, Sheng Yang, Zishuo Zhao, Yu-Kun Lai, Shi-Min Hu; Proceedings o

f the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5875-5884
We present a practical backend for stereo visual SLAM which can simultaneously discover individual rigid bodies and compute their motions in dynamic environments. While recent factor graph based state optimization algorithms have shown their ability to robustly solve SLAM problems by treating dynamic objects as outliers, the dynamic motions are rarely considered. In this paper, we exploit the consensus of 3D motions among the landmarks extracted from the same rigid body for clustering and estimating static and dynamic objects in a unified manner. Specifically, our algorithm builds a noise-aware motion affinity matrix upon landmarks, and uses agglomerative clustering for distinguishing those rigid bodies. Accompanied by a decoupled factor graph optimization for revising their shape and trajectory, we obtain an iterative scheme to update both cluster assignments and motion estimation reciprocally. Evaluations on both synthetic scenes and KITTI demonstrate the capability of our approach, and further experiments considering online efficiency also show the effectiveness of our method for simultaneous tracking of ego-motion and multiple objects.
****

Efficient and Robust Registration on the 3D Special Euclidean Group
Uttaran Bhattacharya, Venu Madhav Govindu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5885-5894
We present a robust, fast and accurate method for registration of 3D scans. Using correspondences, our method optimizes a robust cost function on the intrinsic representation of rigid motions, i.e., the Special Euclidean group SE(3). We exploit the geometric properties of Lie groups as well as the robustness afforded by an iteratively reweighted least squares optimization. We also generalize our approach to a joint multiview method that simultaneously solves for the registration of a set of scans. Our approach significantly outperforms the state-of-the-art robust 3D registration method based on a line process in terms of both speed and accuracy. We show that this line process method is a special case of our principled geometric solution. Finally, we also present scenarios where global registration based on feature correspondences fails but multiview ICP based on our robust motion estimation is successful.
****

Algebraic Characterization of Essential Matrices and Their Averaging in Multiview Settings
Yoni Kasten, Amnon Geifman, Meirav Galun, Ronen Basri; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5895-5903
Essential matrix averaging, i.e., the task of recovering camera locations and orientations in calibrated, multiview settings, is a first step in global approaches to Euclidean structure from motion. A common approach to essential matrix averaging is to separately solve for camera orientations and subsequently for camera positions. This paper presents a novel approach that solves simultaneously for both camera orientations and positions. We offer a complete characterization of the algebraic conditions that enable a unique Euclidean reconstruction of n cameras from a collection of (^n_2) essential matrices. We next use these conditions to formulate essential matrix averaging as a constrained optimization problem, allowing us to recover a consistent set of essential matrices given a (possibly partial) set of measured essential matrices computed independently for pairs of images. We finally use the recovered essential matrices to determine the global positions and orientations of the n cameras. We test our method on common SfM datasets, demonstrating high accuracy while maintaining efficiency and robustness, compared to existing methods.
****

Liquid Warping GAN: A Unified Framework for Human Motion Imitation, Appearance Transfer and Novel View Synthesis
Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, Shenghua Gao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5904-5913
We tackle the human motion imitation, appearance transfer, and novel view synthe

sis within a unified framework, which means that the model once being trained ca n be used to handle all these tasks. The existing task-specific methods mainly u se 2D keypoints (pose) to estimate the human body structure. However, they only expresses the position information with no abilities to characterize the persona lized shape of the individual person and model the limbs rotations. In this pape r, we propose to use a 3D body mesh recovery module to disentangle the pose and shape, which can not only model the joint location and rotation but also charact erize the personalized body shape. To preserve the source information, such as t exture, style, color, and face identity, we propose a Liquid Warping GAN with Li quid Warping Block (LWB) that propagates the source information in both image an d feature spaces, and synthesizes an image with respect to the reference. Specif ically, the source features are extracted by a denoising convolutional auto-enco der for characterizing the source identity well. Furthermore, our proposed metho d is able to support a more flexible warping from multiple sources. In addition, we build a new dataset, namely Impersonator (iPER) dataset, for the evaluation of human motion imitation, appearance transfer, and novel view synthesis. Extens ive experiments demonstrate the effectiveness of our method in several aspects, such as robustness in occlusion case and preserving face identity, shape consist ency and clothes details. All codes and datasets are available on https://svip-l ab.github.io/project/impersonator.html.

****

RelGAN: Multi-Domain Image-to-Image Translation via Relative Attributes

Po-Wei Wu, Yu-Jing Lin, Che-Han Chang, Edward Y. Chang, Shih-Wei Liao; Proce edings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5914-5922

Multi-domain image-to-image translation has gained increasing attention recently . Previous methods take an image and some target attributes as inputs and genera te an output image with the desired attributes. However, such methods have two l imitations. First, these methods assume binary-valued attributes and thus cannot yield satisfactory results for fine-grained control. Second, these methods requ ire specifying the entire set of target attributes, even if most of the attribut es would not be changed. To address these limitations, we propose RelGAN, a new method for multi-domain image-to-image translation. The key idea is to use relat ive attributes, which describes the desired change on selected attributes. Our m ethod is capable of modifying images by changing particular attributes of intere st in a continuous manner while preserving the other attributes. Experimental re sults demonstrate both the quantitative and qualitative effectiveness of our met hod on the tasks of facial attribute transfer and interpolation.

****

Attribute-Driven Spontaneous Motion in Unpaired Image Translation

Ruizheng Wu, Xin Tao, Xiaodong Gu, Xiaoyong Shen, Jiaya Jia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5923-5932

Current image translation methods, albeit effective to produce high-quality resu lts in various applications, still do not consider much geometric transform. We in this paper propose the spontaneous motion estimation module, along with a ref inement part, to learn attribute-driven deformation between source and target do mains. Extensive experiments and visualization demonstrate effectiveness of thes e modules. We achieve promising results in unpaired-image translation tasks, and enable interesting applications based on spontaneous motion.

****

Everybody Dance Now

Caroline Chan, Shiry Ginosar, Tinghui Zhou, Alexei A. Efros; Proceedings of t he IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5933-5 942

This paper presents a simple method for "do as I do" motion transfer: given a so urce video of a person dancing, we can transfer that performance to a novel (ama teur) target after only a few minutes of the target subject performing standard moves. We approach this problem as video-to-video translation using pose as an i ntermediate representation. To transfer the motion, we extract poses from the so

urce subject and apply the learned pose-to-appearance mapping to generate the ta rget subject. We predict two consecutive frames for temporally coherent video re sults and introduce a separate pipeline for realistic face synthesis. Although o ur method is quite simple, it produces surprisingly compelling results (see vide o). This motivates us to also provide a forensics tool for reliable synthetic co ntent detection, which is able to distinguish videos synthesized by our system f rom real data. In addition, we release a first-of-its-kind open-source dataset o f videos that can be legally used for training and motion transfer.
****

Multimodal Style Transfer via Graph Cuts
Yulun Zhang, Chen Fang, Yilin Wang, Zhaowen Wang, Zhe Lin, Yun Fu, Jimei Y ang; Proceedings of the IEEE/CVF International Conference on Computer Vision (IC CV), 2019, pp. 5943-5951
An assumption widely used in recent neural style transfer methods is that image styles can be described by global statics of deep features like Gram or covarian ce matrices. Alternative approaches have represented styles by decomposing them into local pixel or neural patches. Despite the recent progress, most existing m ethods treat the semantic patterns of style image uniformly, resulting unpleasin g results on complex styles. In this paper, we introduce a more flexible and gen eral universal style transfer technique: multimodal style transfer (MST). MST ex plicitly considers the matching of semantic patterns in content and style images . Specifically, the style image features are clustered into sub-style components , which are matched with local content features under a graph cut formulation. A reconstruction network is trained to transfer each sub-style and render the fin al stylized result. We also generalize MST to improve some existing methods. Ext ensive experiments demonstrate the superior effectiveness, robustness, and flexi bility of MST.
****

A Closed-Form Solution to Universal Style Transfer
Ming Lu, Hao Zhao, Anbang Yao, Yurong Chen, Feng Xu, Li Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 59 52-5961
Universal style transfer tries to explicitly minimize the losses in feature spac e, thus it does not require training on any pre-defined styles. It usually uses different layers of VGG network as the encoders and trains several decoders to i nvert the features into images. Therefore, the effect of style transfer is achie ved by feature transform. Although plenty of methods have been proposed, a theor etical analysis of feature transform is still missing. In this paper, we first p ropose a novel interpretation by treating it as the optimal transport problem. T hen, we demonstrate the relations of our formulation with former works like Adap tive Instance Normalization (AdaIN) and Whitening and Coloring Transform (WCT). Finally, we derive a closed-form solution named Optimal Style Transfer (OST) und er our formulation by additionally considering the content loss of Gatys. Compar atively, our solution can preserve better structure and achieve visually pleasin g results. It is simple yet effective and we demonstrate its advantages both qua ntitatively and qualitatively. Besides, we hope our theoretical analysis can ins pire future works in neural style transfer.
****

Progressive Reconstruction of Visual Structure for Image Inpainting
Jingyuan Li, Fengxiang He, Lefei Zhang, Bo Du, Dacheng Tao; Proceedings of t he IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5962-5 971
Inpainting methods aim to restore missing parts of corrupted images and play a c ritical role in many computer vision applications, such as object removal and im age restoration. Although existing methods perform well on images with small hol es, restoring large holes remains elusive. To address this issue, this paper pro poses a Progressive Reconstruction of Visual Structure (PRVS) network that progr essively reconstructs the structures and the associated visual feature. Specific ally, we design a novel Visual Structure Reconstruction (VSR) layer to entangle reconstructions of the visual structure and visual feature, which benefits each

other by sharing parameters. We repeatedly stack four VSR layers in both encodin g and decoding stages of a U-Net like architecture to form the generator of a ge nerative adversarial network (GAN) for restoring images with either small or lar ge holes. We prove the generalization error upper bound of the PRVS network is O (1sqrt(N)), which theoretically guarantees its performance. Extensive empirical evaluations and comparisons on Places2, Paris Street View and CelebA datasets va lidate the strengths of the proposed approach and demonstrate that the model out performs current state-of-the-art methods. The source code package is available at https://github.com/jingyuanli001/PRVS-Image-Inpainting.
****

Variational Adversarial Active Learning
Samarth Sinha, Sayna Ebrahimi, Trevor Darrell; Proceedings of the IEEE/CVF Int ernational Conference on Computer Vision (ICCV), 2019, pp. 5972-5981
Active learning aims to develop label-efficient algorithms by sampling the most representative queries to be labeled by an oracle. We describe a pool-based semi -supervised active learning algorithm that implicitly learns this sampling mecha nism in an adversarial manner. Our method learns a latent space using a variatio nal autoencoder (VAE) and an adversarial network trained to discriminate between unlabeled and labeled data. The mini-max game between the VAE and the adversari al network is played such that while the VAE tries to trick the adversarial netw ork into predicting that all data points are from the labeled pool, the adversar ial network learns how to discriminate between dissimilarities in the latent spa ce. We extensively evaluate our method on various image classification and seman tic segmentation benchmark datasets and establish a new state of the art on CIFA R10/100, Caltech-256, ImageNet, Cityscapes, and BDD100K. Our results demonstrate that our adversarial approach learns an effective low dimensional latent space in large-scale settings and provides for a computationally efficient sampling me thod. Our code is available at https://github.com/sinhasam/vaal.
****

Confidence Regularized Self-Training
Yang Zou, Zhiding Yu, Xiaofeng Liu, B.V.K. Vijaya Kumar, Jinsong Wang; Proce edings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5982-5991
Recent advances in domain adaptation show that deep self-training presents a pow erful means for unsupervised domain adaptation. These methods often involve an i terative process of predicting on target domain and then taking the confident pr edictions as pseudo-labels for retraining. However, since pseudo-labels can be n oisy, self-training can put overconfident label belief on wrong classes, leading to deviated solutions with propagated errors. To address the problem, we propos e a confidence regularized self-training (CRST) framework, formulated as regular ized self-training. Our method treats pseudo-labels as continuous latent variabl es jointly optimized via alternating optimization. We propose two types of confi dence regularization: label regularization (LR) and model regularization (MR). C RST-LR generates soft pseudo-labels while CRST-MR encourages the smoothness on n etwork output. Extensive experiments on image classification and semantic segmen tation show that CRSTs outperform their non-regularized counterpart with state-o f-the-art performance. The code and models of this work are available at https:/ /github.com/yzou2/CRST.
****

Anchor Loss: Modulating Loss Scale Based on Prediction Difficulty
Serim Ryou, Seong-Gyun Jeong, Pietro Perona; Proceedings of the IEEE/CVF Inter national Conference on Computer Vision (ICCV), 2019, pp. 5992-6001
We propose a novel loss function that dynamically re-scales the cross entropy ba sed on prediction difficulty regarding a sample. Deep neural network architectur es in image classification tasks struggle to disambiguate visually similar objec ts. Likewise, in human pose estimation symmetric body parts often confuse the ne twork with assigning indiscriminative scores to them. This is due to the output prediction, in which only the highest confidence label is selected without takin g into consideration a measure of uncertainty. In this work, we define the predi ction difficulty as a relative property coming from the confidence score gap bet

ween positive and negative labels. More precisely, the proposed loss function pe
nalizes the network to avoid the score of a false prediction being significant.
To demonstrate the efficacy of our loss function, we evaluate it on two differen
t domains: image classification and human pose estimation. We find improvements
in both applications by achieving higher accuracy compared to the baseline metho
ds.
****

Local Aggregation for Unsupervised Learning of Visual Embeddings
Chengxu Zhuang, Alex Lin Zhai, Daniel Yamins; Proceedings of the IEEE/CVF Inte
rnational Conference on Computer Vision (ICCV), 2019, pp. 6002-6012
Unsupervised approaches to learning in neural networks are of substantial intere
st for furthering artificial intelligence, both because they would enable the tr
aining of networks without the need for large numbers of expensive annotations,
and because they would be better models of the kind of general-purpose learning
deployed by humans. However, unsupervised networks have long lagged behind the p
erformance of their supervised counterparts, especially in the domain of large-s
cale visual recognition. Recent developments in training deep convolutional embe
ddings to maximize non-parametric instance separation and clustering objectives
have shown promise in closing this gap. Here, we describe a method that trains a
n embedding function to maximize a metric of local aggregation, causing similar
data instances to move together in the embedding space, while allowing dissimila
r instances to separate. This aggregation metric is dynamic, allowing soft clust
ers of different scales to emerge. We evaluate our procedure on several large-sc
ale visual recognition datasets, achieving state-of-the-art unsupervised transfe
r learning performance on object recognition in ImageNet, scene recognition in P
laces 205, and object detection in PASCAL VOC.
****

PR Product: A Substitute for Inner Product in Neural Networks
Zhennan Wang, Wenbin Zou, Chen Xu; Proceedings of the IEEE/CVF International C
onference on Computer Vision (ICCV), 2019, pp. 6013-6022
In this paper, we analyze the inner product of weight vector w and data vector x
in neural networks from the perspective of vector orthogonal decomposition and
prove that the direction gradient of w decreases with the angle between them clo
se to 0 or p. We propose the Projection and Rejection Product (PR Product) to ma
ke the direction gradient of w independent of the angle and consistently larger
than the one in standard inner product while keeping the forward propagation ide
ntical. As a reliable substitute for standard inner product, the PR Product can
be applied into many existing deep learning modules, so we develop the PR Produc
t version of fully connected layer, convolutional layer and LSTM layer. In stati
c image classification, the experiments on CIFAR10 and CIFAR100 datasets demonst
rate that the PR Product can robustly enhance the ability of various state-of-th
e-art classification networks. On the task of image captioning, even without any
bells and whistles, our PR Product version of captioning model can compete or o
utperform the state-of-the-art models on MS COCO dataset. Code has been made ava
ilable at: https://github.com/wzn0828/PR_Product.
****

CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Fea
tures
Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, Young
joon Yoo; Proceedings of the IEEE/CVF International Conference on Computer Visio
n (ICCV), 2019, pp. 6023-6032
Regional dropout strategies have been proposed to enhance performance of convolu
tional neural network classifiers. They have proved to be effective for guiding
the model to attend on less discriminative parts of objects (e.g. leg as opposed
to head of a person), thereby letting the network generalize better and have be
tter object localization capabilities. On the other hand, current methods for re
gional dropout removes informative pixels on training images by overlaying a pat
ch of either black pixels or random noise. Such removal is not desirable because
it suffers from information loss causing inefficiency in training. We therefore
propose the CutMix augmentation strategy: patches are cut and pasted among trai

ning images where the ground truth labels are also mixed proportionally to the a rea of the patches. By making efficient use of training pixels and retaining the regularization effect of regional dropout, CutMix consistently outperforms stat e-of-the-art augmentation strategies on CIFAR and ImageNet classification tasks, as well as on ImageNet weakly-supervised localization task. Moreover, unlike pr evious augmentation methods, our CutMix-trained ImageNet classifier, when used a s a pretrained model, results in consistent performance gain in Pascal detection and MS-COCO image captioning benchmarks. We also show that CutMix can improve t he model robustness against input corruptions and its out-of distribution detect ion performance.

****

Towards Interpretable Object Detection by Unfolding Latent Structures
Tianfu Wu, Xi Song; Proceedings of the IEEE/CVF International Conference on Com puter Vision (ICCV), 2019, pp. 6033-6043

This paper first proposes a method of formulating model interpretability in visu al understanding tasks based on the idea of unfolding latent structures. It then presents a case study in object detection using popular two-stage region-based convolutional network (i.e., R-CNN) detection systems. The proposed method focus es on weakly-supervised extractive rationale generation, that is learning to unf old latent discriminative part configurations of object instances automatically and simultaneously in detection without using any supervision for part configura tions. It utilizes a top-down hierarchical and compositional grammar model embed ded in a directed acyclic AND-OR Graph (AOG) to explore and unfold the space of latent part configurations of regions of interest (RoIs). It presents an AOGPars ing operator that seamlessly integrates with the RoIPooling/RoIAlign operator wi dely used in R-CNN and is trained end-to-end. In object detection, a bounding bo x is interpreted by the best parse tree derived from the AOG on-the-fly, which i s treated as the qualitatively extractive rationale generated for interpreting d etection. In experiments, Faster R-CNN is used to test the proposed method on th e PASCAL VOC 2007 and the COCO 2017 object detection datasets. The experimental results show that the proposed method can compute promising latent structures wi thout hurting the performance. The code and pretrained models are available at h ttps://github.com/iVMCL/iRCNN.

****

Scaling Object Detection by Transferring Classification Weights
Jason Kuen, Federico Perazzi, Zhe Lin, Jianming Zhang, Yap-Peng Tan; Proceed ings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, p p. 6044-6053

Large scale object detection datasets are constantly increasing their size in te rms of the number of classes and annotations count. Yet, the number of object-le vel categories annotated in detection datasets is an order of magnitude smaller than image-level classification labels. State-of-the art object detection models are trained in a supervised fashion and this limits the number of object classe s they can detect. In this paper, we propose a novel weight transfer network (WT N) to effectively and efficiently transfer knowledge from classification network 's weights to detection network's weights to allow detection of novel classes wi thout box supervision. We first introduce input and feature normalization scheme s to curb the under-fitting during training of a vanilla WTN. We then propose au toencoder-WTN (AE-WTN) which uses reconstruction loss to preserve classification network's information over all classes in the target latent space to ensure gen eralization to novel classes. Compared to vanilla WTN, AE-WTN obtains absolute p erformance gains of 6% on two Open Images evaluation sets with 500 seen and 57 n ovel classes respectively, and 25% on a Visual Genome evaluation set with 200 no vel classes.

****

Scale-Aware Trident Networks for Object Detection
Yanghao Li, Yuntao Chen, Naiyan Wang, Zhaoxiang Zhang; Proceedings of the IEE E/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6054-6063

Scale variation is one of the key challenges in object detection. In this work, we first present a controlled experiment to investigate the effect of receptive

fields for scale variation in object detection. Based on the findings from the exploration experiments, we propose a novel Trident Network (TridentNet) aiming to generate scale-specific feature maps with a uniform representational power. We construct a parallel multi-branch architecture in which each branch shares the same transformation parameters but with different receptive fields. Then, we adopt a scale-aware training scheme to specialize each branch by sampling object instances of proper scales for training. As a bonus, a fast approximation version of TridentNet could achieve significant improvements without any additional parameters and computational cost compared with the vanilla detector. On the COCO dataset, our TridentNet with ResNet-101 backbone achieves state-of-the-art single-model results of 48.4 mAP. Codes are available at https://git.io/fj5vR.
****

Object-Aware Instance Labeling for Weakly Supervised Object Detection
Satoshi Kosugi, Toshihiko Yamasaki, Kiyoharu Aizawa; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6064-6072
Weakly supervised object detection (WSOD), where a detector is trained with only image-level annotations, is attracting more and more attention. As a method to obtain a well-performing detector, the detector and the instance labels are updated iteratively. In this study, for more efficient iterative updating, we focus on the instance labeling problem, a problem of which label should be annotated to each region based on the last localization result. Instead of simply labeling the top-scoring region and its highly overlapping regions as positive and others as negative, we propose more effective instance labeling methods as follows. First, to solve the problem that regions covering only some parts of the object tend to be labeled as positive, we find regions covering the whole object focusing on the context classification loss. Second, considering the situation where the other objects contained in the image can be labeled as negative, we impose a spatial restriction on regions labeled as negative. Using these instance labeling methods, we train the detector on the PASCAL VOC 2007 and 2012 and obtain significantly improved results compared with other state-of-the-art approaches.
****

Generative Modeling for Small-Data Object Detection
Lanlan Liu, Michael Muelly, Jia Deng, Tomas Pfister, Li-Jia Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6073-6081
This paper explores object detection in the small data regime, where only a limited number of annotated bounding boxes are available due to data rarity and annotation expense. This is a common challenge today with machine learning being applied to many new tasks where obtaining training data is more challenging, e.g. in medical images with rare diseases that doctors sometimes only see once in their life-time. In this work we explore this problem from a generative modeling perspective by learning to generate new images with associated bounding boxes, and using these for training an object detector. We show that simply training previously proposed generative models does not yield satisfactory performance due to them optimizing for image realism rather than object detection accuracy. To this end we develop a new model with a novel unrolling mechanism that jointly optimizes the generative model and a detector such that the generated images improve the performance of the detector. We show this method outperforms the state of the art on two challenging datasets, disease detection and small data pedestrian detection, improving the average precision on NIH Chest X-ray by a relative 20% and localization accuracy by a relative 50%.
****

Transductive Learning for Zero-Shot Object Detection
Shafin Rahman, Salman Khan, Nick Barnes; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6082-6091
Zero-shot object detection (ZSD) is a relatively unexplored research problem as compared to the conventional zero-shot recognition task. ZSD aims to detect previously unseen objects during inference. Existing ZSD works suffer from two critical issues: (a) large domain-shift between the source (seen) and target (unseen) domains since the two distributions are highly mismatched. (b) the learned mode

l is biased against unseen classes, therefore in generalized ZSD settings, where both seen and unseen objects co-occur during inference, the learned model tends to misclassify unseen to seen categories. This brings up an important question: How effectively can a transductive setting address the aforementioned problems? To the best of our knowledge, we are the first to propose a transductive zero-shot object detection approach that convincingly reduces the domain-shift and model-bias against unseen classes. Our approach is based on a self-learning mechanism that uses a novel hybrid pseudo-labeling technique. It progressively updates learned model parameters by associating unlabeled data samples to their corresponding classes. During this process, our technique makes sure that knowledge that was previously acquired on the source domain is not forgotten. We report significant 'relative' improvements of 34.9% and 77.1% in terms of mAP and recall rates over the previous best inductive models on MSCOCO dataset.
****

Self-Training and Adversarial Background Regularization for Unsupervised Domain Adaptive One-Stage Object Detection
Seunghyeon Kim, Jaehoon Choi, Taekyung Kim, Changick Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6092-6101
Deep learning-based object detectors have shown remarkable improvements. However, supervised learning-based methods perform poorly when the train data and the test data have different distributions. To address the issue, domain adaptation transfers knowledge from the label-sufficient domain (source domain) to the label-scarce domain (target domain). Self-training is one of the powerful ways to achieve domain adaptation since it helps class-wise domain adaptation. Unfortunately, a naive approach that utilizes pseudo-labels as ground-truth degenerates the performance due to incorrect pseudo-labels. In this paper, we introduce a weak self-training (WST) method and adversarial background score regularization (BSR) for domain adaptive one-stage object detection. WST diminishes the adverse effects of inaccurate pseudo-labels to stabilize the learning procedure. BSR helps the network extract discriminative features for target backgrounds to reduce the domain shift. Two components are complementary to each other as BSR enhances discrimination between foregrounds and backgrounds, whereas WST strengthen class-wise discrimination. Experimental results show that our approach effectively improves the performance of the one-stage object detection in unsupervised domain adaptation setting.
****

Memory-Based Neighbourhood Embedding for Visual Recognition
Suichan Li, Dapeng Chen, Bin Liu, Nenghai Yu, Rui Zhao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6102-6111
Learning discriminative image feature embeddings is of great importance to visual recognition. To achieve better feature embeddings, most current methods focus on designing different network structures or loss functions, and the estimated feature embeddings are usually only related to the input images. In this paper, we propose Memory-based Neighbourhood Embedding (MNE) to enhance a general CNN feature by considering its neighbourhood. The method aims to solve two critical problems, i.e., how to acquire more relevant neighbours in the network training and how to aggregate the neighbourhood information for a more discriminative embedding. We first augment an episodic memory module into the network, which can provide more relevant neighbours for both training and testing. Then the neighbours are organized in a tree graph with the target instance as the root node. The neighbourhood information is gradually aggregated to the root node in a bottom-up manner, and aggregation weights are supervised by the class relationships between the nodes. We apply MNE on image search and few shot learning tasks. Extensive ablation studies demonstrate the effectiveness of each component, and our method significantly outperforms the state-of-the-art approaches.
****

Self-Similarity Grouping: A Simple Unsupervised Cross Domain Adaptation Approach for Person Re-Identification
Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, Thomas S. Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (IC

CV), 2019, pp. 6112-6121
Domain adaptation in person re-identification (re-ID) has always been a challeng ing task. In this work, we explore how to harness the similar natural characteri stics existing in the samples from the target domain for learning to conduct per son re-ID in an unsupervised manner. Concretely, we propose a Self-similarity Gr ouping (SSG) approach, which exploits the potential similarity (from the global body to local parts) of unlabeled samples to build multiple clusters from differ ent views automatically. These independent clusters are then assigned with label s, which serve as the pseudo identities to supervise the training process. We re peatedly and alternatively conduct such a grouping and training process until th e model is stable. Despite the apparent simplify, our SSG outperforms the state- of-the-arts by more than 4.6% (DukeMTMC-Market1501) and 4.4% (Market1501-DukeMTM C) in mAP, respectively. Upon our SSG, we further introduce a clustering-guided semisupervised approach named SSG ++ to conduct the one-shot domain adaption in an open set setting (i.e. the number of independent identities from the target d omain is unknown). Without spending much effort on labeling, our SSG ++ can furt her promote the mAP upon SSG by 10.7% and 6.9%, respectively. Our Code is availa ble at: https://github.com/OasisYang/SSG .
****

Deep Reinforcement Active Learning for Human-in-the-Loop Person Re-Identificatio n
Zimo Liu, Jingya Wang, Shaogang Gong, Huchuan Lu, Dacheng Tao; Proceedings o f the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 612 2-6131
Most existing person re-identification(Re-ID) approaches achieve superior result s based on the assumption that a large amount of pre-labelled data is usually av ailable and can be put into training phrase all at once. However, this assumptio n is not applicable to most real-world deployment of the Re-ID task. In this wor k, we propose an alternative reinforcement learning based human-in-the-loop mode l which releases the restriction of pre-labelling and keeps model upgrading with progressively collected data. The goal is to minimize human annotation efforts while maximizing Re-ID performance. It works in an iteratively updating framewor k by refining the RL policy and CNN parameters alternately. In particular, we fo rmulate a Deep Reinforcement Active Learning (DRAL) method to guide an agent (a model in a reinforcement learning process) in selecting training samples on-the- fly by a human user/annotator. The reinforcement learning reward is the uncertai nty value of each human selected sample. A binary feedback (positive or negative) labelled by the human annotator is used to select the samples of which are use d to fine-tune a pre-trained CNN Re-ID model. Extensive experiments demonstrate the superiority of our DRAL method for deep reinforcement learning based human-i n-the-loop person Re-ID when compared to existing unsupervised and transfer lear ning models as well as active learning models.
****

A Dual-Path Model With Adaptive Attention for Vehicle Re-Identification
Pirazh Khorramshahi, Amit Kumar, Neehar Peri, Sai Saketh Rambhatla, Jun-Chen g Chen, Rama Chellappa; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6132-6141
In recent years, attention models have been extensively used for person and vehi cle re-identification. Most re-identification methods are designed to focus atte ntion on key-point locations. However, depending on the orientation, the contrib ution of each key-point varies. In this paper, we present a novel dual-path adap tive attention model for vehicle re-identification (AAVER). The global appearanc e path captures macroscopic vehicle features while the orientation conditioned p art appearance path learns to capture localized discriminative features by focus ing attention on the most informative key-points. Through extensive experimentat ion, we show that the proposed AAVER method is able to accurately re-identify ve hicles in unconstrained scenarios, yielding state of the art results on the chal lenging dataset VeRi-776. As a byproduct, the proposed system is also able to ac curately predict vehicle key-points and shows an improvement of more than 7% ove r state of the art. The code for key-point estimation model is available at http

s://github.com/Pirazh/Vehicle_Key_ Point_Orientation_Estimation
****

Bayesian Loss for Crowd Count Estimation With Point Supervision
Zhiheng Ma, Xing Wei, Xiaopeng Hong, Yihong Gong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6142-6151
In crowd counting datasets, each person is annotated by a point, which is usually the center of the head. And the task is to estimate the total count in a crowd scene. Most of the state-of-the-art methods are based on density map estimation, which convert the sparse point annotations into a "ground truth" density map through a Gaussian kernel, and then use it as the learning target to train a density map estimator. However, such a "ground-truth" density map is imperfect due to occlusions, perspective effects, variations in object shapes, etc. On the contrary, we propose Bayesian loss, a novel loss function which constructs a density contribution probability model from the point annotations. Instead of constraining the value at every pixel in the density map, the proposed training loss adopts a more reliable supervision on the count expectation at each annotated point. Without bells and whistles, the loss function makes substantial improvements over the baseline loss on all tested datasets. Moreover, our proposed loss function equipped with a standard backbone network, without using any external detectors or multi-scale architectures, plays favourably against the state of the arts. Our method outperforms previous best approaches by a large margin on the latest and largest UCF-QNRF dataset.
****

Learning Spatial Awareness to Improve Crowd Counting
Zhi-Qi Cheng, Jun-Xiu Li, Qi Dai, Xiao Wu, Alexander G. Hauptmann; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6152-6161
The aim of crowd counting is to estimate the number of people in images by leveraging the annotation of center positions for pedestrians' heads. Promising progresses have been made with the prevalence of deep Convolutional Neural Networks. Existing methods widely employ the Euclidean distance (i.e., L_2 loss) to optimize the model, which, however, has two main drawbacks: (1) the loss has difficulty in learning the spatial awareness (i.e., the position of head) since it struggles to retain the high-frequency variation in the density map, and (2) the loss is highly sensitive to various noises in crowd counting, such as the zero-mean noise, head size changes, and occlusions. Although the Maximum Excess over SubArrays (MESA) loss has been previously proposed by [??] to address the above issues by finding the rectangular subregion whose predicted density map has the maximum difference from the ground truth, it cannot be solved by gradient descent, thus can hardly be integrated into the deep learning framework. In this paper, we present a novel architecture called SPatial Awareness Network (SPANet) to incorporate spatial context for crowd counting. The Maximum Excess over Pixels (MEP) loss is proposed to achieve this by finding the pixel-level subregion with high discrepancy to the ground truth. To this end, we devise a weakly supervised learning scheme to generate such region with a multi-branch architecture. The proposed framework can be integrated into existing deep crowd counting methods and is end-to-end trainable. Extensive experiments on four challenging benchmarks show that our method can significantly improve the performance of baselines. More remarkably, our approach outperforms the state-of-the-art methods on all benchmark datasets.
****

GradNet: Gradient-Guided Network for Visual Object Tracking
Peixia Li, Boyu Chen, Wanli Ouyang, Dong Wang, Xiaoyun Yang, Huchuan Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6162-6171
The fully-convolutional siamese network based on template matching has shown great potentials in visual tracking. During testing, the template is fixed with the initial target feature and the performance totally relies on the general matching ability of the siamese network. However, this manner cannot capture the temporal variations of targets or background clutter. In this work, we propose a nove

l gradient-guided network to exploit the discriminative information in gradients and update the template in the siamese network through feed-forward and backward operations. To be specific, the algorithm can utilize the information from the gradient to update the template in the current frame. In addition, a template generalization training method is proposed to better use gradient information and avoid overfitting. To our knowledge, this work is the first attempt to exploit the information in the gradient for template update in siamese-based trackers. Extensive experiments on recent benchmarks demonstrate that our method achieves better performance than other state-of-the-art trackers.
****

FAMNet: Joint Learning of Feature, Affinity and Multi-Dimensional Assignment for Online Multiple Object Tracking
Peng Chu, Haibin Ling; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6172-6181
Data association-based multiple object tracking (MOT) involves multiple separated modules processed or optimized differently, which results in complex method design and requires non-trivial tuning of parameters. In this paper, we present an end-to-end model, named FAMNet, where Feature extraction, Affinity estimation and Multi-dimensional assignment are refined in a single network. All layers in FAMNet are designed differentiable thus can be optimized jointly to learn the discriminative features and higher-order affinity model for robust MOT, which is supervised by the loss directly from the assignment ground truth. In addition, we integrate single object tracking technique and a dedicated target management scheme into the FAMNet-based tracking system to further recover false negatives and inhibit noisy target candidates generated by the external detector. The proposed method is evaluated on a diverse set of benchmarks including MOT2015, MOT2017, KITTI-Car and UA-DETRAC, and achieves promising performance on all of them in comparison with state-of-the-arts.
****

Learning Discriminative Model Prediction for Tracking
Goutam Bhat, Martin Danelljan, Luc Van Gool, Radu Timofte; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6182-6191
The current strive towards end-to-end trainable computer vision systems imposes major challenges for the task of visual tracking. In contrast to most other vision problems, tracking requires the learning of a robust target-specific appearance model online, during the inference stage. To be end-to-end trainable, the online learning of the target model thus needs to be embedded in the tracking architecture itself. Due to the imposed challenges, the popular Siamese paradigm simply predicts a target feature template, while ignoring the background appearance information during inference. Consequently, the predicted model possesses limited target-background discriminability. We develop an end-to-end tracking architecture, capable of fully exploiting both target and background appearance information for target model prediction. Our architecture is derived from a discriminative learning loss by designing a dedicated optimization process that is capable of predicting a powerful model in only a few iterations. Furthermore, our approach is able to learn key aspects of the discriminative loss itself. The proposed tracker sets a new state-of-the-art on 6 tracking benchmarks, achieving an EAO score of 0.440 on VOT2018, while running at over 40 FPS. The code and models are available at https://github.com/visionml/pytracking.
****

DynamoNet: Dynamic Action and Motion Network
Ali Diba, Vivek Sharma, Luc Van Gool, Rainer Stiefelhagen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6192-6201
In this paper, we are interested in self-supervised learning the motion cues in videos using dynamic motion filters for a better motion representation to finally boost human action recognition in particular. Thus far, the vision community has focused on spatio-temporal approaches using standard filters, rather we here propose dynamic filters that adaptively learn the video-specific internal motion

representation by predicting the short-term future frames. We name this new motion representation, as dynamic motion representation (DMR) and is embedded inside of 3D convolutional network as a new layer, which captures the visual appearance and motion dynamics throughout entire video clip via end-to-end network learning. Simultaneously, we utilize these motion representation to enrich video classification. We have designed the frame prediction task as an auxiliary task to empower the classification problem. With these overall objectives, to this end, we introduce a novel unified spatio-temporal 3D-CNN architecture (DynamoNet) that jointly optimizes the video classification and learning motion representation by predicting future frames as a multi-task learning problem. We conduct experiments on challenging human action datasets: Kinetics 400, UCF101, HMDB51. The experiments using the proposed DynamoNet show promising results on all the datasets.

****

SlowFast Networks for Video Recognition

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, Kaiming He; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6202-6211

We present SlowFast networks for video recognition. Our model involves (i) a Slow pathway, operating at low frame rate, to capture spatial semantics, and (ii) a Fast pathway, operating at high frame rate, to capture motion at fine temporal resolution. The Fast pathway can be made very lightweight by reducing its channel capacity, yet can learn useful temporal information for video recognition. Our models achieve strong performance for both action classification and detection in video, and large improvements are pin-pointed as contributions by our SlowFast concept. We report state-of-the-art accuracy on major video recognition benchmarks, Kinetics, Charades and AVA. Code has been made available at: https://github.com/facebookresearch/SlowFast.

****

Generative Multi-View Human Action Recognition

Lichen Wang, Zhengming Ding, Zhiqiang Tao, Yunyu Liu, Yun Fu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6212-6221

Multi-view action recognition targets to integrate complementary information from different views to improve classification performance. It is a challenging task due to the distinct gap between heterogeneous feature domains. Moreover, most existing methods neglect to consider the incomplete multi-view data, which limits their potential compatibility in real-world applications. In this work, we propose a Generative Multi-View Action Recognition (GMVAR) framework to address the challenges above. The adversarial generative network is leveraged to generate one view conditioning on the other view, which fully explores the latent connections in both intra-view and cross-view aspects. Our approach enhances the model robustness by employing adversarial training, and naturally handles the incomplete view case by imputing the missing data. Moreover, an effective View Correlation Discovery Network (VCDN) is proposed to further fuse the multi-view information in a higher-level label space. Extensive experiments demonstrate the effectiveness of our proposed approach by comparing with state-of-the-art algorithms.

****

Multi-Agent Reinforcement Learning Based Frame Sampling for Effective Untrimmed Video Recognition

Wenhao Wu, Dongliang He, Xiao Tan, Shifeng Chen, Shilei Wen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6222-6231

Video Recognition has drawn great research interest and great progress has been made. A suitable frame sampling strategy can improve the accuracy and efficiency of recognition. However, mainstream solutions generally adopt hand-crafted frame sampling strategies for recognition. It could degrade the performance, especially in untrimmed videos, due to the variation of frame-level saliency. To this end, we concentrate on improving untrimmed video classification via developing a learning-based frame sampling strategy. We intuitively formulate the frame sampling procedure as multiple parallel Markov decision processes, each of which aims

at picking out a frame/clip by gradually adjusting an initial sampling. Then we propose to solve the problems with multi-agent reinforcement learning (MARL). Our MARL framework is composed of a novel RNN-based context-aware observation network which jointly models context information among nearby agents and historical states of a specific agent, a policy network which generates the probability distribution over a predefined action space at each step and a classification network for reward calculation as well as final recognition. Extensive experimental results show that our MARL-based scheme remarkably outperforms hand-crafted strategies with various 2D and 3D baseline methods. Our single RGB model achieves a comparable performance of ActivityNet v1.3 champion submission with multi-modal multi-model fusion and new state-of-the-art results on YouTube Birds and YouTube Cars.

****

SCSampler: Sampling Salient Clips From Video for Efficient Action Recognition
Bruno Korbar, Du Tran, Lorenzo Torresani; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6232-6242

While many action recognition datasets consist of collections of brief, trimmed videos each containing a relevant action, videos in the real-world (e.g., on YouTube) exhibit very different properties: they are often several minutes long, where brief relevant clips are often interleaved with segments of extended duration containing little change. Applying densely an action recognition system to every temporal clip within such videos is prohibitively expensive. Furthermore, as we show in our experiments, this results in suboptimal recognition accuracy as informative predictions from relevant clips are outnumbered by meaningless classification outputs over long uninformative sections of the video. In this paper we introduce a lightweight "clip-sampling" model that can efficiently identify the most salient temporal clips within a long video. We demonstrate that the computational cost of action recognition on untrimmed videos can be dramatically reduced by invoking recognition only on these most salient clips. Furthermore, we show that this yields significant gains in recognition accuracy compared to analysis of all clips or randomly selected clips. On Sports1M, our clip sampling scheme elevates the accuracy of an already state-of-the-art action classifier by 7% and reduces by more than 15 times its computational cost.

****

Weakly Supervised Energy-Based Learning for Action Segmentation
Jun Li, Peng Lei, Sinisa Todorovic; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6243-6251

This paper is about labeling video frames with action classes under weak supervision in training, where we have access to a temporal ordering of actions, but their start and end frames in training videos are unknown. Following prior work, we use an HMM grounded on a Gated Recurrent Unit (GRU) for frame labeling. Our key contribution is a new constrained discriminative forward loss (CDFL) that we use for training the HMM and GRU under weak supervision. While prior work typically estimates the loss on a single, inferred video segmentation, our CDFL discriminates between the energy of all valid and invalid frame labelings of a training video. A valid frame labeling satisfies the ground-truth temporal ordering of actions, whereas an invalid one violates the ground truth. We specify an efficient recursive algorithm for computing the CDFL in terms of the logadd function of the segmentation energy. Our evaluation on action segmentation and alignment gives superior results to those of the state of the art on the benchmark Breakfast Action, Hollywood Extended, and 50Salads datasets.

****

What Would You Expect? Anticipating Egocentric Actions With Rolling-Unrolling LSTMs and Modality Attention
Antonino Furnari, Giovanni Maria Farinella; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6252-6261

Egocentric action anticipation consists in understanding which objects the camera wearer will interact with in the near future and which actions they will perform. We tackle the problem proposing an architecture able to anticipate actions at multiple temporal scales using two LSTMs to 1) summarize the past, and 2) form

ulate predictions about the future. The input video is processed considering thr ee complimentary modalities: appearance (RGB), motion (optical flow) and objects (object-based features). Modality-specific predictions are fused using a novel Modality ATTention (MATT) mechanism which learns to weigh modalities in an adapt ive fashion. Extensive evaluations on two large-scale benchmark datasets show th at our method outperforms prior art by up to +7% on the challenging EPIC-Kitchen s dataset including more than 2500 actions, and generalizes to EGTEA Gaze+. Our approach is also shown to generalize to the tasks of early action recognition an d action recognition. Our method is ranked first in the public leaderboard of th e EPIC-Kitchens egocentric action anticipation challenge 2019. Please see the pr oject web page for code and additional details: http://iplab.dmi.unict.it/rulstm .

****

PIE: A Large-Scale Dataset and Models for Pedestrian Intention Estimation and Tr ajectory Prediction

Amir Rasouli, Iuliia Kotseruba, Toni Kunic, John K. Tsotsos; Proceedings of t he IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6262-6 271

Pedestrian behavior anticipation is a key challenge in the design of assistive a nd autonomous driving systems suitable for urban environments. An intelligent sy stem should be able to understand the intentions or underlying motives of pedest rians and to predict their forthcoming actions. To date, only a few public datas ets were proposed for the purpose of studying pedestrian behavior prediction in the context of intelligent driving. To this end, we propose a novel large-scale dataset designed for pedestrian intention estimation (PIE). We conducted a large -scale human experiment to establish human reference data for pedestrian intenti on in traffic scenes. We propose models for estimating pedestrian crossing inten tion and predicting their future trajectory. Our intention estimation model achi eves 79% accuracy and our trajectory prediction algorithm outperforms state-of-t he-art by 26% on the proposed dataset. We further show that combining pedestrian intention with observed motion improves trajectory prediction. The dataset and models are available at http://data.nvision2.eecs.yorku.ca/PIE_dataset/.

****

STGAT: Modeling Spatial-Temporal Interactions for Human Trajectory Prediction

Yingfan Huang, Huikun Bi, Zhaoxin Li, Tianlu Mao, Zhaoqi Wang; Proceedings o f the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 627 2-6281

Human trajectory prediction is challenging and critical in various applications (e.g., autonomous vehicles and social robots). Because of the continuity and for esight of the pedestrian movements, the moving pedestrians in crowded spaces wil l consider both spatial and temporal interactions to avoid future collisions. Ho wever, most of the existing methods ignore the temporal correlations of interact ions with other pedestrians involved in a scene. In this work, we propose a Spat ial-Temporal Graph Attention network (STGAT), based on a sequence-to-sequence ar chitecture to predict future trajectories of pedestrians. Besides the spatial in teractions captured by the graph attention mechanism at each time-step, we adopt an extra LSTM to encode the temporal correlations of interactions. Through comp arisons with state-of-the-art methods, our model achieves superior performance o n two publicly available crowd datasets (ETH and UCY) and produces more "sociall y" plausible trajectories for pedestrians.

****

Learning Motion in Feature Space: Locally-Consistent Deformable Convolution Netw orks for Fine-Grained Action Detection

Khoi-Nguyen C. Mac, Dhiraj Joshi, Raymond A. Yeh, Jinjun Xiong, Rogerio S. F eris, Minh N. Do; Proceedings of the IEEE/CVF International Conference on Compu ter Vision (ICCV), 2019, pp. 6282-6291

Fine-grained action detection is an important task with numerous applications in robotics and human-computer interaction. Existing methods typically utilize a t wo-stage approach including extraction of local spatio-temporal features followe d by temporal modeling to capture long-term dependencies. While most recent pape

rs have focused on the latter (long-temporal modeling), here, we focus on produc ing features capable of modeling fine-grained motion more efficiently. We propos e a novel locally-consistent deformable convolution, which utilizes the change i n receptive fields and enforces a local coherency constraint to capture motion i nformation effectively. Our model jointly learns spatio-temporal features (inste ad of using independent spatial and temporal streams). The temporal component is learned from the feature space instead of pixel space, e.g. optical flow. The p roduced features can be flexibly used in conjunction with other long-temporal mo deling networks, e.g. ST-CNN, DilatedTCN, and ED-TCN. Overall, our proposed appr oach robustly outperforms the original long-temporal models on two fine-grained action datasets: 50 Salads and GTEA, achieving F1 scores of 80.22% and 75.39% re spectively.

****

Dual Attention Matching for Audio-Visual Event Localization

Yu Wu, Linchao Zhu, Yan Yan, Yi Yang; Proceedings of the IEEE/CVF Internation al Conference on Computer Vision (ICCV), 2019, pp. 6292-6300

In this paper, we investigate the audio-visual event localization problem. This task is to localize a visible and audible event in a video. Previous methods fir st divide a video into short segments, and then fuse visual and acoustic feature s at the segment level. The duration of these segments is usually short, making the visual and acoustic feature of each segment possibly not well aligned. Direc t concatenation of the two features at the segment level can be vulnerable to a minor temporal misalignment of the two signals. We propose a Dual Attention Matc hing (DAM) module to cover a longer video duration for better high-level event i nformation modeling, while the local temporal information is attained by the glo bal cross-check mechanism. Our premise is that one should watch the whole video to understand the high-level event, while shorter segments should be checked in detail for localization. Specifically, the global feature of one modality querie s the local feature in the other modality in a bi-directional way. With temporal co-occurrence encoded between auditory and visual signals, DAM can be readily a pplied in various audio-visual event localization tasks, e.g., cross-modality lo calization, supervised event localization. Experiments on the AVE dataset show o ur method outperforms the state-of-the-art by a large margin.

****

Uncertainty-Aware Audiovisual Activity Recognition Using Deep Bayesian Variation al Inference

Mahesh Subedar, Ranganath Krishnan, Paulo Lopez Meyer, Omesh Tickoo, Jonatha n Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6301-6310

Deep neural networks (DNNs) provide state-of-the-art results for a multitude of applications, but the approaches using DNNs for multimodal audiovisual applicati ons do not consider predictive uncertainty associated with individual modalities . Bayesian deep learning methods provide principled confidence and quantify pred ictive uncertainty. Our contribution in this work is to propose an uncertainty a ware multimodal Bayesian fusion framework for activity recognition. We demonstra te a novel approach that combines deterministic and variational layers to scale Bayesian DNNs to deeper architectures. Our experiments using in- and out-of-dist ribution samples selected from a subset of Moments-in-Time (MiT) dataset show a more reliable confidence measure as compared to the non-Bayesian baseline and th e Monte Carlo dropout (MC dropout) approximate Bayesian inference. We also demon strate the uncertainty estimates obtained from the proposed framework can identi fy out-of-distribution data on the UCF101 and MiT datasets. In the multimodal se tting, the proposed framework improved precision-recall AUC by 10.2% on the subs et of MiT dataset as compared to non-Bayesian baseline.

****

Non-Local Recurrent Neural Memory for Supervised Sequence Modeling

Canmiao Fu, Wenjie Pei, Qiong Cao, Chaopeng Zhang, Yong Zhao, Xiaoyong Shen , Yu-Wing Tai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6311-6320

Typical methods for supervised sequence modeling are built upon the recurrent ne

ural networks to capture temporal dependencies. One potential limitation of thes e methods is that they only model explicitly information interactions between ad jacent time steps in a sequence, hence the high-order interactions between nonad jacent time steps are not fully exploited. It greatly limits the capability of m odeling the long-range temporal dependencies since one-order interactions cannot be maintained for a long term due to information dilution and gradient vanishin g. To tackle this limitation, we propose the Non-local Recurrent Neural Memory (NRNM) for supervised sequence modeling, which performs non-local operations to l earn full-order interactions within a sliding temporal block and models the glob al interactions between blocks in a gated recurrent manner. Consequently, our mo del is able to capture the long-range dependencies. Besides, the latent high-lev el features contained in high-order interactions can be distilled by our model. We demonstrate the merits of our NRNM approach on two different tasks: action re cognition and sentiment analysis.

****

Temporal Attentive Alignment for Large-Scale Video Domain Adaptation

Min-Hung Chen, Zsolt Kira, Ghassan AlRegib, Jaekwon Yoo, Ruxin Chen, Jian Z heng; Proceedings of the IEEE/CVF International Conference on Computer Vision (I CCV), 2019, pp. 6321-6330

Although various image-based domain adaptation (DA) techniques have been propose d in recent years, domain shift in videos is still not well-explored. Most previ ous works only evaluate performance on small-scale datasets which are saturated. Therefore, we first propose two large-scale video DA datasets with much larger domain discrepancy: UCF-HMDB_full and Kinetics-Gameplay. Second, we investigate different DA integration methods for videos, and show that simultaneously aligni ng and learning temporal dynamics achieves effective alignment even without soph isticated DA methods. Finally, we propose Temporal Attentive Adversarial Adaptat ion Network (TA3N), which explicitly attends to the temporal dynamics using doma in discrepancy for more effective domain alignment, achieving state-of-the-art p erformance on four video DA datasets (e.g. 7.9% accuracy gain over "Source only" from 73.9% to 81.8% on "HMDB --> UCF", and 10.3% gain on "Kinetics --> Gameplay "). The code and data are released at http://github.com/cmhungsteve/TA3N.

****

Action Assessment by Joint Relation Graphs

Jia-Hui Pan, Jibin Gao, Wei-Shi Zheng; Proceedings of the IEEE/CVF Internation al Conference on Computer Vision (ICCV), 2019, pp. 6331-6340

We present a new model to assess the performance of actions from videos, through graph-based joint relation modelling. Previous works mainly focused on the whol e scene including the performer's body and background, yet they ignored the deta iled joint interactions. This is insufficient for fine-grained, accurate action assessment, because the action quality of each joint is dependent of its neighbo uring joints. Therefore, we propose to learn the detailed joint motion based on the joint relations. We build trainable Joint Relation Graphs, and analyze joint motion on them. We propose two novel modules, the Joint Commonality Module and the Joint Difference Module, for joint motion learning. The Joint Commonality Mo dule models the general motion for certain body parts, and the Joint Difference Module models the motion differences within body parts. We evaluate our method o n six public Olympic actions for performance assessment. Our method outperforms previous approaches (+0.0912) and the whole-scene analysis (+0.0623) in the Spea rman's Rank Correlation. We also demonstrate our model's ability to interpret th e action assessment process.

****

Unsupervised Procedure Learning via Joint Dynamic Summarization

Ehsan Elhamifar, Zwe Naing; Proceedings of the IEEE/CVF International Conferenc e on Computer Vision (ICCV), 2019, pp. 6341-6350

We address the problem of unsupervised procedure learning from unconstrained ins tructional videos. Our goal is to produce a summary of the procedure key-steps a nd their ordering needed to perform a given task, as well as localization of the key-steps in videos. We develop a collaborative sequential subset selection fra mework, where we build a dynamic model on videos by learning states and transiti

ons between them, where states correspond to different subactivities, including background and procedure steps. To extract procedure key-steps, we develop an optimization framework that finds a sequence of a small number of states that well represents all videos and is compatible with the state transition model. Given that our proposed optimization is non-convex and NP-hard, we develop a fast greedy algorithm whose complexity is linear in the length of the videos and the number of states of the dynamic model, hence, scales to large datasets. Under appropriate conditions on the transition model, our proposed formulation is approximately submodular, hence, comes with performance guarantees. We also present ProceL, a new multimodal dataset of 47.3 hours of videos and their transcripts from diverse tasks, for procedure learning evaluation. By extensive experiments, we show that our framework significantly improves the state of the art performance.
****

ViSiL: Fine-Grained Spatio-Temporal Video Similarity Learning
Giorgos Kordopatis-Zilos, Symeon Papadopoulos, Ioannis Patras, Ioannis Kompatsiaris; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6351-6360
In this paper we introduce ViSiL, a Video Similarity Learning architecture that considers fine-grained Spatio-Temporal relations between pairs of videos -- such relations are typically lost in previous video retrieval approaches that embed the whole frame or even the whole video into a vector descriptor before the similarity estimation. By contrast, our Convolutional Neural Network (CNN)-based approach is trained to calculate video-to-video similarity from refined frame-to-frame similarity matrices, so as to consider both intra- and inter-frame relations. In the proposed method, pairwise frame similarity is estimated by applying Tensor Dot (TD) followed by Chamfer Similarity (CS) on regional CNN frame features - this avoids feature aggregation before the similarity calculation between frames. Subsequently, the similarity matrix between all video frames is fed to a four-layer CNN, and then summarized using Chamfer Similarity (CS) into a video-to-video similarity score -- this avoids feature aggregation before the similarity calculation between videos and captures the temporal similarity patterns between matching frame sequences. We train the proposed network using a triplet loss scheme and evaluate it on five public benchmark datasets on four different video retrieval problems where we demonstrate large improvements in comparison to the state of the art. The implementation of ViSiL is publicly available.
****

Unsupervised Learning of Landmarks by Descriptor Vector Exchange
James Thewlis, Samuel Albanie, Hakan Bilen, Andrea Vedaldi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6361-6371
Equivariance to random image transformations is an effective method to learn landmarks of object categories, such as the eyes and the nose in faces, without manual supervision. However, this method does not explicitly guarantee that the learned landmarks are consistent with changes between different instances of the same object, such as different facial identities. In this paper, we develop a new perspective on the equivariance approach by noting that dense landmark detectors can be interpreted as local image descriptors equipped with invariance to intra-category variations. We then propose a direct method to enforce such an invariance in the standard equivariant loss. We do so by exchanging descriptor vectors between images of different object instances prior to matching them geometrically. In this manner, the same vectors must work regardless of the specific object identity considered. We use this approach to learn vectors that can simultaneously be interpreted as local descriptors and dense landmarks, combining the advantages of both. Experiments on standard benchmarks show that this approach can match, and in some cases surpass state-of-the-art performance amongst existing methods that learn landmarks without supervision. Code is available at www.robots.ox.ac.uk/ vgg/research/DVE/.
****

Learning Compositional Representations for Few-Shot Recognition
Pavel Tokmakov, Yu-Xiong Wang, Martial Hebert; Proceedings of the IEEE/CVF Int

ernational Conference on Computer Vision (ICCV), 2019, pp. 6372-6381
One of the key limitations of modern deep learning approaches lies in the amount of data required to train them. Humans, by contrast, can learn to recognize novel categories from just a few examples. Instrumental to this rapid learning ability is the compositional structure of concept representations in the human brain --- something that deep learning models are lacking. In this work, we make a step towards bridging this gap between human and machine learning by introducing a simple regularization technique that allows the learned representation to be decomposable into parts. Our method uses category-level attribute annotations to disentangle the feature space of a network into subspaces corresponding to the attributes. These attributes can be either purely visual, like object parts, or more abstract, like openness and symmetry. We demonstrate the value of compositional representations on three datasets: CUB-200-2011, SUN397, and ImageNet, and show that they require fewer examples to learn classifiers for novel categories.
****

Spectral Regularization for Combating Mode Collapse in GANs
Kanglin Liu, Wenming Tang, Fei Zhou, Guoping Qiu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6382-6390
Despite excellent progress in recent years, mode collapse remains a major unsolved problem in generative adversarial networks (GANs). In this paper, we present spectral regularization for GANs (SR-GANs), a new and robust method for combating the mode collapse problem in GANs. Theoretical analysis shows that the optimal solution to the discriminator has a strong relationship to the spectral distributions of the weight matrix. Therefore, we monitor the spectral distribution in the discriminator of spectral normalized GANs (SN-GANs), and discover a phenomenon which we refer to as spectral collapse, where a large number of singular values of the weight matrices drop dramatically when mode collapse occurs. We show that there are strong evidence linking mode collapse to spectral collapse; and based on this link, we set out to tackle spectral collapse as a surrogate of mode collapse. We have developed a spectral regularization method where we compensate the spectral distributions of the weight matrices to prevent them from collapsing, which in turn successfully prevents mode collapse in GANs. We provide theoretical explanations for why SR-GANs are more stable and can provide better performances than SN-GANs. We also present extensive experimental results and analysis to show that SR-GANs not only always outperform SN-GANs but also always succeed in combating mode collapse where SN-GANs fail.
****

Scaling and Benchmarking Self-Supervised Visual Representation Learning
Priya Goyal, Dhruv Mahajan, Abhinav Gupta, Ishan Misra; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6391-6400
Self-supervised learning aims to learn representations from the data itself without explicit manual supervision. Existing efforts ignore a crucial aspect of self-supervised learning - the ability to scale to large amount of data because self-supervision requires no manual labels. In this work, we revisit this principle and scale two popular self-supervised approaches to 100 million images. We show that by scaling on various axes (including data size and problem 'hardness'), one can largely match or even exceed the performance of supervised pre-training on a variety of tasks such as object detection, surface normal estimation (3D) and visual navigation using reinforcement learning. Scaling these methods also provides many interesting insights into the limitations of current self-supervised techniques and evaluations. We conclude that current self-supervised methods are not 'hard' enough to take full advantage of large scale data and do not seem to learn effective high level semantic representations. We also introduce an extensive benchmark across 9 different datasets and tasks. We believe that such a benchmark along with comparable evaluation settings is necessary to make meaningful progress. Code is at: https://github.com/facebookresearch/fair_self_supervision_benchmark.
****

Learning an Effective Equivariant 3D Descriptor Without Supervision
Riccardo Spezialetti, Samuele Salti, Luigi Di Stefano; Proceedings of the IEEE

/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6401-6410
Establishing correspondences between 3D shapes is a fundamental task in 3D Compu ter Vision, typically ad- dressed by matching local descriptors. Recently, a few at- tempts at applying the deep learning paradigm to the task have shown promis ing results. Yet, the only explored way to learn rotation invariant descriptors has been to feed neural networks with highly engineered and invariant representa - tions provided by existing hand-crafted descriptors, a path that goes in the o pposite direction of end-to-end learning from raw data so successfully deployed for 2D images. In this paper, we explore the benefits of taking a step back in t he direction of end-to-end learning of 3D descrip- tors by disentangling the cre ation of a robust and distinctive rotation equivariant representation, which can be learned from unoriented input data, and the definition of a good canonical o rientation, required only at test time to obtain an invariant descriptor. To thi s end, we leverage two re- cent innovations: spherical convolutional neural netw orks to learn an equivariant descriptor and plane folding de- coders to learn wi thout supervision. The effectiveness of the proposed approach is experimentally validated by out- performing hand-crafted and learned descriptors on a stan- dar d benchmark.
****

KPConv: Flexible and Deformable Convolution for Point Clouds
Hugues Thomas, Charles R. Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, Fr ancois Goulette, Leonidas J. Guibas; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6411-6420
We present Kernel Point Convolution (KPConv), a new design of point convolution, i.e. that operates on point clouds without any intermediate representation. The convolution weights of KPConv are located in Euclidean space by kernel points, and applied to the input points close to them. Its capacity to use any number of kernel points gives KPConv more flexibility than fixed grid convolutions. Furth ermore, these locations are continuous in space and can be learned by the networ k. Therefore, KPConv can be extended to deformable convolutions that learn to ad apt kernel points to local geometry. Thanks to a regular subsampling strategy, K PConv is also efficient and robust to varying densities. Whether they use deform able KPConv for complex tasks, or rigid KPconv for simpler tasks, our networks o utperform state-of-the-art classification and segmentation approaches on several datasets. We also offer ablation studies and visualizations to provide understa nding of what has been learned by KPConv and to validate the descriptive power o f deformable KPConv.
****

Neural Inter-Frame Compression for Video Coding
Abdelaziz Djelouah, Joaquim Campos, Simone Schaub-Meyer, Christopher Schroers ; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) , 2019, pp. 6421-6429
While there are many deep learning based approaches for single image compression , the field of end-to-end learned video coding has remained much less explored. Therefore, in this work we present an inter-frame compression approach for neura l video coding that can seamlessly build up on different existing neural image c odecs. Our end-to-end solution performs temporal prediction by optical flow base d motion compensation in pixel space. The key insight is that we can increase bo th decoding efficiency and reconstruction quality by encoding the required infor mation into a latent representation that directly decodes into motion and blendi ng coefficients. In order to account for remaining prediction errors, residual i nformation between the original image and the interpolated frame is needed. We p ropose to compute residuals directly in latent space instead of in pixel space a s this allows to reuse the same image compression network for both key frames an d intermediate frames. Our extended evaluation on different datasets and resolut ions shows that the rate-distortion performance of our approach is competitive w ith existing state-of-the-art codecs.
****

Task2Vec: Task Embedding for Meta-Learning
Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhran

su Maji, Charless C. Fowlkes, Stefano Soatto, Pietro Perona; Proceedings of t
he IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6430-6
439
We introduce a method to generate vectorial representations of visual classifica
tion tasks which can be used to reason about the nature of those tasks and their
 relations. Given a dataset with ground-truth labels and a loss function, we pro
cess images through a "probe network" and compute an embedding based on estimate
s of the Fisher information matrix associated with the probe network parameters.
 This provides a fixed-dimensional embedding of the task that is independent of
details such as the number of classes and requires no understanding of the class
 label semantics. We demonstrate that this embedding is capable of predicting ta
sk similarities that match our intuition about semantic and taxonomic relations
between different visual tasks. We demonstrate the practical value of this frame
work for the meta-task of selecting a pre-trained feature extractor for a novel
task. We present a simple meta-learning framework for learning a metric on embed
dings that is capable of predicting which feature extractors will perform well o
n which task. Selecting a feature extractor with task embedding yields performan
ce close to the best available feature extractor, with substantially less comput
ational effort than exhaustively training and evaluating all available models.
****

Deep Clustering by Gaussian Mixture Variational Autoencoders With Graph Embeddin
g
Linxiao Yang, Ngai-Man Cheung, Jiaying Li, Jun Fang; Proceedings of the IEEE/
CVF International Conference on Computer Vision (ICCV), 2019, pp. 6440-6449
We propose DGG: D eep clustering via a G aussian-mixture variational autoencod
er (VAE) with G raph embedding. To facilitate clustering, we apply Gaussian mix
ture model (GMM) as the prior in VAE. To handle data with complex spread, we app
ly graph embedding. Our idea is that graph information which captures local data
 structures is an excellent complement to deep GMM. Combining them facilitates t
he network to learn powerful representations that follow global model and local
structural constraints. Therefore, our method unifies model-based and similarity
-based approaches for clustering. To combine graph embedding with probabilistic
deep GMM, we propose a novel stochastic extension of graph embedding: we treat s
amples as nodes on a graph and minimize the weighted distance between their post
erior distributions. We apply Jenson-Shannon divergence as the distance. We comb
ine the divergence minimization with the log-likelihood maximization of the deep
 GMM. We derive formulations to obtain an unified objective that enables simulta
neous deep representation learning and clustering. Our experimental results show
 that our proposed DGG outperforms recent deep Gaussian mixture methods (model-b
ased) and deep spectral clustering (similarity-based). Our results highlight adv
antages of combining model-based and similarity-based clustering as proposed in
this work. Our code is published here: https://github.com/dodoyang0929/DGG.git
****

SoftTriple Loss: Deep Metric Learning Without Triplet Sampling
Qi Qian, Lei Shang, Baigui Sun, Juhua Hu, Hao Li, Rong Jin; Proceedings of
the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6450-
6458
Distance metric learning (DML) is to learn the embeddings where examples from th
e same class are closer than examples from different classes. It can be cast as
an optimization problem with triplet constraints. Due to the vast number of trip
let constraints, a sampling strategy is essential for DML. With the tremendous s
uccess of deep learning in classifications, it has been applied for DML. When le
arning embeddings with deep neural networks (DNNs), only a mini-batch of data is
 available at each iteration. The set of triplet constraints has to be sampled w
ithin the mini-batch. Since a mini-batch cannot capture the neighbors in the ori
ginal set well, it makes the learned embeddings sub-optimal. On the contrary, op
timizing SoftMax loss, which is a classification loss, with DNN shows a superior
 performance in certain DML tasks. It inspires us to investigate the formulation
 of SoftMax. Our analysis shows that SoftMax loss is equivalent to a smoothed tr
iplet loss where each class has a single center. In real-world data, one class c

an contain several local clusters rather than a single one, e.g., birds of diffe rent poses. Therefore, we propose the SoftTriple loss to extend the SoftMax loss with multiple centers for each class. Compared with conventional deep metric le arning algorithms, optimizing SoftTriple loss can learn the embeddings without t he sampling phase by mildly increasing the size of the last fully connected laye r. Experiments on the benchmark fine-grained data sets demonstrate the effective ness of the proposed loss function.

****

A Weakly Supervised Fine Label Classifier Enhanced by Coarse Supervision

Fariborz Taherkhani, Hadi Kazemi, Ali Dabouei, Jeremy Dawson, Nasser M. Nasr abadi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6459-6468

Objects are usually organized in a hierarchical structure in which each coarse c ategory (e.g., big cat) corresponds to a superclass of several fine categories (e.g., cheetah, leopard). The objects grouped within the same coarse category, bu t in different fine categories, usually share a set of global visual features; h owever, these objects have distinctive local properties that characterize them a t a fine level. This paper addresses the challenge of fine image classification in a weakly supervised fashion, whereby a subset of images is tagged by fine lab els, while the remaining are tagged by coarse labels. We propose a new deep mode l that leverages coarse images to improve the classification performance of fine images within the coarse category. Our model is an end to end framework consist ing of a Convolutional Neural Network (CNN) which uses both fine and coarse imag es to tune its parameters. The CNN outputs are then fanned out into two separate branches such that the first branch uses a supervised low rank self expressive layer to project the CNN outputs to the low rank subspaces to capture the global structures for the coarse classification, while the other branch uses a supervi sed sparse self expressive layer to project them to the sparse subspaces to capt ure the local structures for the fine classification. Our deep model uses coarse images in conjunction with fine images to jointly explore the low rank and spar se subspaces by sharing the parameters during the training which causes the data points obtained by the CNN to be well-projected to both sparse and low rank sub spaces for classification.

****

Gaussian Affinity for Max-Margin Class Imbalanced Learning

Munawar Hayat, Salman Khan, Syed Waqas Zamir, Jianbing Shen, Ling Shao; Proc eedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019 , pp. 6469-6479

Real-world object classes appear in imbalanced ratios. This poses a significant challenge for classifiers which get biased towards frequent classes. We hypothes ize that improving the generalization capability of a classifier should improve learning on imbalanced datasets. Here, we introduce the first hybrid loss functi on that jointly performs classification and clustering in a single formulation. Our approach is based on an `affinity measure' in Euclidean space that leads to the following benefits: (1) direct enforcement of maximum margin constraints on classification boundaries, (2) a tractable way to ensure uniformly spaced and eq uidistant cluster centers, (3) flexibility to learn multiple class prototypes to support diversity and discriminability in feature space. Our extensive experime nts demonstrate the significant performance improvements on visual classificatio n and verification tasks on multiple imbalanced datasets. The proposed loss can easily be plugged in any deep architecture as a differentiable block and demonst rates robustness against different levels of data imbalance and corrupted labels .

****

AttPool: Towards Hierarchical Feature Representation in Graph Convolutional Netw orks via Attention Mechanism

Jingjia Huang, Zhangheng Li, Nannan Li, Shan Liu, Ge Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6480-6489

Graph convolutional networks (GCNs) are potentially short of the ability to lear n hierarchical representation for graph embedding, which holds them back in the

graph classification task. Here, we propose AttPool, which is a novel graph pool ing module based on attention mechanism, to remedy the problem. It is able to se lect nodes that are significant for graph representation adaptively, and generat e hierarchical features via aggregating the attention-weighted information in no des. Additionally, we devise a hierarchical prediction architecture to sufficien tly leverage the hierarchical representation and facilitate the model learning. The AttPool module together with the entire training structure can be integrated into existing GCNs, and is trained in an end-to-end fashion conveniently. The e xperimental results on several graph-classification benchmark datasets with vari ous scales demonstrate the effectiveness of our method.
****

Deep Metric Learning With Tuplet Margin Loss
Baosheng Yu, Dacheng Tao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6490-6499
Deep metric learning, in which the loss function plays a key role, has proven to be extremely useful in visual recognition tasks. However, existing deep metric learning loss functions such as contrastive loss and triplet loss usually rely o n delicately selected samples (pairs or triplets) for fast convergence. In this paper, we propose a new deep metric learning loss function, tuplet margin loss, using randomly selected samples from each mini-batch. Specifically, the proposed tuplet margin loss implicitly up-weights hard samples and down-weights easy sam ples, while a slack margin in angular space is introduced to mitigate the proble m of overfitting on the hardest sample. Furthermore, we address the problem of i ntra-pair variation by disentangling class-specific information to improve the g eneralizability of tuplet margin loss. Experimental results on three widely used deep metric learning datasets, CARS196, CUB200-2011, and Stanford Online Produc ts, demonstrate significant improvements over existing deep metric learning meth ods.
****

Normalized Wasserstein for Mixture Distributions With Applications in Adversaria l Learning and Domain Adaptation
Yogesh Balaji, Rama Chellappa, Soheil Feizi; Proceedings of the IEEE/CVF Inter national Conference on Computer Vision (ICCV), 2019, pp. 6500-6508
Understanding proper distance measures between distributions is at the core of s everal learning tasks such as generative models, domain adaptation, clustering, etc. In this work, we focus on mixture distributions that arise naturally in sev eral application domains where the data contains different sub-populations. For mixture distributions, established distance measures such as the Wasserstein dis tance do not take into account imbalanced mixture proportions. Thus, even if two mixture distributions have identical mixture components but different mixture p roportions, the Wasserstein distance between them will be large. This often lead s to undesired results in distance-based learning methods for mixture distributi ons. In this paper, we resolve this issue by introducing the Normalized Wasserst ein measure. The key idea is to introduce mixture proportions as optimization va riables, effectively normalizing mixture proportions in the Wasserstein formulat ion. Using the proposed normalized Wasserstein measure leads to significant perf ormance gains for mixture distributions with imbalanced mixture proportions comp ared to the vanilla Wasserstein distance. We demonstrate the effectiveness of th e proposed measure in GANs, domain adaptation and adversarial clustering in seve ral benchmark datasets.
****

Fast and Practical Neural Architecture Search
Jiequan Cui, Pengguang Chen, Ruiyu Li, Shu Liu, Xiaoyong Shen, Jiaya Jia; P roceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2 019, pp. 6509-6518
In this paper, we propose a fast and practical neural architecture search (FPNAS) framework for automatic network design. FPNAS aims to discover extremely effic ient networks with less than 300M FLOPs. Different from previous NAS methods, ou r approach searches for the whole network architecture to guarantee block divers ity instead of stacking a set of similar blocks repeatedly. We model the search

process as a bi-level optimization problem and propose an approximation solution . On CIFAR-10, our approach is capable of design networks with comparable perfor mance to state-of-the-arts while using orders of magnitude less computational re source with only 20 GPU hours. Experimental results on ImageNet and ADE20K datas ets further demonstrate transferability of the searched networks.
****

Symmetric Graph Convolutional Autoencoder for Unsupervised Graph Representation Learning
Jiwoong Park, Minsik Lee, Hyung Jin Chang, Kyuewang Lee, Jin Young Choi; Pro ceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 201 9, pp. 6519-6528
We propose a symmetric graph convolutional autoencoder which produces a low-dime nsional latent representation from a graph. In contrast to the existing graph au toencoders with asymmetric decoder parts, the proposed autoencoder has a newly d esigned decoder which builds a completely symmetric autoencoder form. For the re construction of node features, the decoder is designed based on Laplacian sharpe ning as the counterpart of Laplacian smoothing of the encoder, which allows util izing the graph structure in the whole processes of the proposed autoencoder arc hitecture. In order to prevent the numerical instability of the network caused b y the Laplacian sharpening introduction, we further propose a new numerically st able form of the Laplacian sharpening by incorporating the signed graphs. In add ition, a new cost function which finds a latent representation and a latent affi nity matrix simultaneously is devised to boost the performance of image clusteri ng tasks. The experimental results on clustering, link prediction and visualizat ion tasks strongly support that the proposed model is stable and outperforms var ious state-of-the-art algorithms.
****

Deep Elastic Networks With Model Selection for Multi-Task Learning
Chanho Ahn, Eunwoo Kim, Songhwai Oh; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6529-6538
In this work, we consider the problem of instance-wise dynamic network model sel ection for multi-task learning. To this end, we propose an efficient approach to exploit a compact but accurate model in a backbone architecture for each instan ce of all tasks. The proposed method consists of an estimator and a selector. Th e estimator is based on a backbone architecture and structured hierarchically. I t can produce multiple different network models of different configurations in a hierarchical structure. The selector chooses a model dynamically from a pool of candidate models given an input instance. The selector is a relatively small-si ze network consisting of a few layers, which estimates a probability distributio n over the candidate models when an input instance of a task is given. Both esti mator and selector are jointly trained in a unified learning framework in conjun ction with a sampling-based learning strategy, without additional computation st eps. We demonstrate the proposed approach for several image classification tasks compared to existing approaches performing model selection or learning multiple tasks. Experimental results show that our approach gives not only outstanding p erformance compared to other competitors but also the versatility to perform ins tance-wise model selection for multiple tasks.
****

Metric Learning With HORDE: High-Order Regularizer for Deep Embeddings
Pierre Jacob, David Picard, Aymeric Histace, Edouard Klein; Proceedings of th e IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6539-65 48
Learning an effective similarity measure between image representations is key to the success of recent advances in visual search tasks (e.g. verification or zer o-shot learning). Although the metric learning part is well addressed, this metr ic is usually computed over the average of the extracted deep features. This rep resentation is then trained to be discriminative. However, these deep features t end to be scattered across the feature space. Consequently, the representations are not robust to outliers, object occlusions, background variations, etc. In th is paper, we tackle this scattering problem with a distribution-aware regulariza

tion named HORDE. This regularizer enforces visually-close images to have deep features with the same distribution which are well localized in the feature space. We provide a theoretical analysis supporting this regularization effect. We also show the effectiveness of our approach by obtaining state-of-the-art results on 4 well-known datasets (Cub-200-2011, Cars-196, Stanford Online Products and Inshop Clothes Retrieval).

****

Adversarial Learning With Margin-Based Triplet Embedding Regularization

Yaoyao Zhong, Weihong Deng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6549-6558

The Deep neural networks (DNNs) have achieved great success on a variety of computer vision tasks, however, they are highly vulnerable to adversarial attacks. To address this problem, we propose to improve the local smoothness of the representation space, by integrating a margin-based triplet embedding regularization term into the classification objective, so that the obtained models learn to resist adversarial examples. The regularization term consists of two steps optimizations which find potential perturbations and punish them by a large margin in an iterative way. Experimental results on MNIST, CASIA-WebFace, VGGFace2 and MS-Celeb-1M reveal that our approach increases the robustness of the network against both feature and label adversarial attacks in simple object classification and deep face recognition.

****

Simultaneous Multi-View Instance Detection With Learned Geometric Soft-Constraints

Ahmed Samy Nassar, Sebastien Lefevre, Jan Dirk Wegner; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6559-6568

We propose to jointly learn multi-view geometry and warping between views of the same object instances for robust cross-view object detection. What makes multi-view object instance detection difficult are strong changes in viewpoint, lighting conditions, high similarity of neighbouring objects, and strong variability in scale. By turning object detection and instance re-identification in different views into a joint learning task, we are able to incorporate both image appearance and geometric soft constraints into a single, multi-view detection process that is learnable end-to-end. We validate our method on a new, large data set of street-level panoramas of urban objects and show superior performance compared to various baselines. Our contribution is threefold: a large-scale, publicly available data set for multi-view instance detection and re-identification; an annotation tool custom-tailored for multi-view instance detection; and a novel, holistic multi-view instance detection and re-identification method that jointly models geometry and appearance across views.

****

CenterNet: Keypoint Triplets for Object Detection

Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, Qi Tian; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6569-6578

In object detection, keypoint-based approaches often experience the drawback of a large number of incorrect object bounding boxes, arguably due to the lack of an additional assessment inside cropped regions. This paper presents an efficient solution that explores the visual patterns within individual cropped regions with minimal costs. We build our framework upon a representative one-stage keypoint-based detector named CornerNet. Our approach, named CenterNet, detects each object as a triplet, rather than a pair, of keypoints, which improves both precision and recall. Accordingly, we design two customized modules, cascade corner pooling, and center pooling, that enrich information collected by both the top-left and bottom-right corners and provide more recognizable information from the central regions. On the MS-COCO dataset, CenterNet achieves an AP of 47.0 %, outperforming all existing one-stage detectors by at least 4.9%. Furthermore, with a faster inference speed than the top-ranked two-stage detectors, CenterNet demonstrates a comparable performance to these detectors. Code is available at https://github.com/Duankaiwen/CenterNet.

****
Online Hyper-Parameter Learning for Auto-Augmentation Strategy
Chen Lin, Minghao Guo, Chuming Li, Xin Yuan, Wei Wu, Junjie Yan, Dahua Lin, Wanli Ouyang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6579-6588
Data augmentation is critical to the success of modern deep learning techniques. In this paper, we propose Online Hyper-parameter Learning for Auto-Augmentation (OHL-Auto-Aug), an economical solution that learns the augmentation policy distribution along with network training. Unlike previous methods on auto-augmentation that search augmentation strategies in an offline manner, our method formulates the augmentation policy as a parameterized probability distribution, thus allowing its parameters to be optimized jointly with network parameters. Our proposed OHL-Auto-Aug eliminates the need of re-training and dramatically reduces the cost of the overall search process, while establishes significantly accuracy improvements over baseline models. On both CIFAR-10 and ImageNet, our method achieves remarkable on search accuracy, 60x faster on CIFAR-10 and 24x faster on ImageNet, while maintaining competitive accuracies.
****
DANet: Divergent Activation for Weakly Supervised Object Localization
Haolan Xue, Chang Liu, Fang Wan, Jianbin Jiao, Xiangyang Ji, Qixiang Ye; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6589-6598
Weakly supervised object localization remains a challenge when learning object localization models from image category labels. Optimizing image classification tends to activate object parts and ignore the full object extent, while expanding object parts into full object extent could deteriorate the performance of image classification. In this paper, we propose a divergent activation (DA) approach, and target at learning complementary and discriminative visual patterns for image classification and weakly supervised object localization from the perspective of discrepancy. To this end, we design hierarchical divergent activation (HDA), which leverages the semantic discrepancy to spread feature activation, implicitly. We also propose discrepant divergent activation (DDA), which pursues object extent by learning mutually exclusive visual patterns, explicitly. Deep networks implemented with HDA and DDA, referred to as DANets, diverge and fuse discrepant yet discriminative features for image classification and object localization in an end-to-end manner. Experiments validate that DANets advance the performance of object localization while maintaining high performance of image classification on CUB-200 and ILSVRC datasets
****
Selective Sparse Sampling for Fine-Grained Image Recognition
Yao Ding, Yanzhao Zhou, Yi Zhu, Qixiang Ye, Jianbin Jiao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6599-6608
Fine-grained recognition poses the unique challenge of capturing subtle inter-class differences under considerable intra-class variances (e.g., beaks for bird species). Conventional approaches crop local regions and learn detailed representation from those regions, but suffer from the fixed number of parts and missing of surrounding context. In this paper, we propose a simple yet effective framework, called Selective Sparse Sampling, to capture diverse and fine-grained details. The framework is implemented using Convolutional Neural Networks, referred to as Selective Sparse Sampling Networks (S3Ns). With image-level supervision, S3Ns collect peaks, i.e., local maximums, from class response maps to estimate informative, receptive fields and learn a set of sparse attention for capturing fine-detailed visual evidence as well as preserving context. The evidence is selectively sampled to extract discriminative and complementary features, which significantly enrich the learned representation and guide the network to discover more subtle cues. Extensive experiments and ablation studies show that the proposed method consistently outperforms the state-of-the-art methods on challenging benchmarks including CUB-200-2011, FGVC-Aircraft, and Stanford Cars.
****

Dynamic Anchor Feature Selection for Single-Shot Object Detection
Shuai Li, Lingxiao Yang, Jianqiang Huang, Xian-Sheng Hua, Lei Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6609-6618
The design of anchors is critical to the performance of one-stage detectors. Recently, the anchor refinement module (ARM) has been proposed to adjust the initialization of default anchors, providing the detector a better anchor reference. However, this module brings another problem: all pixels at a feature map have the same receptive field while the anchors associated with each pixel have different positions and sizes. This discordance may lead to a less effective detector. In this paper, we present a dynamic feature selection operation to select new pixels in a feature map for each refined anchor received from the ARM. The pixels are selected based on the new anchor position and size so that the receptive filed of these pixels can fit the anchor areas well, which makes the detector, especially the regression part, much easier to optimize. Furthermore, to enhance the representation ability of selected feature pixels, we design a bidirectional feature fusion module by combining features from early and deep layers. Extensive experiments on both PASCAL VOC and COCO demonstrate the effectiveness of our dynamic anchor feature selection (DAFS) operation. For the case of high IoU threshold, our DAFS can improve the mAP by a large margin.
****
Incremental Learning Using Conditional Adversarial Networks
Ye Xiang, Ying Fu, Pan Ji, Hua Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6619-6628
Incremental learning using Deep Neural Networks (DNNs) suffers from catastrophic forgetting. Existing methods mitigate it by either storing old image examples or only updating a few fully connected layers of DNNs, which, however, requires large memory footprints or hurts the plasticity of models. In this paper, we propose a new incremental learning strategy based on conditional adversarial networks. Our new strategy allows us to use memory-efficient statistical information to store old knowledge, and fine-tune both convolutional layers and fully connected layers to consolidate new knowledge. Specifically, we propose a model consisting of three parts, i.e., a base sub-net, a generator, and a discriminator. The base sub-net works as a feature extractor which can be pre-trained on large scale datasets and shared across multiple image recognition tasks. The generator conditioned on labeled embeddings aims to construct pseudo-examples with the same distribution as the old data. The discriminator combines real-examples from new data and pseudo-examples generated from the old data distribution to learn representation for both old and new classes. Through adversarial training of the discriminator and generator, we accomplish the multiple continuous incremental learning. Comparison with the state-of-the-arts on public CIFAR-100 and CUB-200 datasets shows that our method achieves the best accuracies on both old and new classes while requiring relatively less memory storage.
****
Bilateral Adversarial Training: Towards Fast Training of More Robust Models Against Adversarial Attacks
Jianyu Wang, Haichao Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6629-6638
In this paper, we study fast training of adversarially robust models. From the analyses of the state-of-the-art defense method, i.e., the multi-step adversarial training [??], we hypothesize that the gradient magnitude links to the model robustness. Motivated by this, we propose to perturb both the image and the label during training, which we call Bilateral Adversarial Training (BAT). To generate the adversarial label, we derive an closed-form heuristic solution. To generate the adversarial image, we use one-step targeted attack with the target label being the most confusing class. In the experiment, we first show that random start and the most confusing target attack effectively prevent the label leaking and gradient masking problem. Then coupled with the adversarial label part, our model significantly improves the state-of-the-art results. For example, against PGD100 white-box attack with cross-entropy loss, on CIFAR10, we achieve 63.7% versus

47.2%; on SVHN, we achieve 59.1% versus 42.1%. At last, the experiment on the very (computationally) challenging ImageNet dataset further demonstrates the effectiveness of our fast method.
****

View Confusion Feature Learning for Person Re-Identification
Fangyi Liu, Lei Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6639-6648
Person re-identification is an important task in video surveillance that aims to associate people across camera views at different locations and time. View variability is always a challenging problem seriously degrading person re-identification performance. Most of the existing methods either focus on how to learn view invariant feature or how to combine viewwise features. In this paper, we mainly focus on how to learn view-independent features by getting rid of view specific information through a view confusion learning mechanism. Specifically, we propose an end-to-end trainable framework, called View Confusion Feature Learning (VCFL), for person Re-ID across cameras. To the best of our knowledge, VCFL is originally proposed to learn view-independent identity-wise features, and it's a kind of combination of view-generic and view-specific methods. Furthermore, we extract sift-guided features by using bag-of-words model to help supervise the training of deep networks and enhance the view invariance of features. In experiments, our approach is validated on three benchmark datasets including CUHK01, CUHK03, and MARKET1501, which show the superiority of the proposed method over several state-of-the-art approaches.
****

Auto-FPN: Automatic Network Architecture Adaptation for Object Detection Beyond Classification
Hang Xu, Lewei Yao, Wei Zhang, Xiaodan Liang, Zhenguo Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6649-6658
Abstract Neural architecture search (NAS) has shown great potential in automating the manual process of designing a good CNN architecture for image classification. In this paper, we study NAS for object detection, a core computer vision task that classifies and localizes object instances in an image. Existing works focus on transferring the searched architecture from classification task (ImageNet) to the detector backbone, while the rest of the architecture of the detector remains unchanged. However, this pipeline is not task-specific or data-oriented network search which cannot guarantee optimal adaptation to any dataset. Therefore, we propose an architecture search framework named Auto-FPN specifically designed for detection beyond simply searching a classification backbone. Specifically, we propose two auto search modules for detection: Auto-fusion to search a better fusion of the multi-level features; Auto-head to search a better structure for classification and bounding-box(bbox) regression. Instead of searching for one repeatable cell structure, we relax the constraint and allow different cells. The search space of both modules covers many popular designs of detectors and allows efficient gradient-based architecture search with resource constraint (2 days for COCO on 8 GPU cards). Extensive experiments on Pascal VOC, COCO, BDD, VisualGenome and ADE demonstrate the effectiveness of the proposed method, e.g. achieving around 5% improvement than FPN in terms of mAP while requiring around 50% fewer parameters on the searched modules.
****

PARN: Position-Aware Relation Networks for Few-Shot Learning
Ziyang Wu, Yuwei Li, Lihua Guo, Kui Jia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6659-6667
Few-shot learning presents a challenge that a classifier must quickly adapt to new classes that do not appear in the training set, given only a few labeled examples of each new class. This paper proposes a position-aware relation network (PARN) to learn a more flexible and robust metric ability for few-shot learning. Relation networks (RNs), a kind of architectures for relational reasoning, can acquire a deep metric ability for images by just being designed as a simple convolutional neural network (CNN)[23]. However, due to the inherent local connectivit

y of CNN, the CNN-based relation network (RN) can be sensitive to the spatial po
sition relationship of semantic objects in two compared images. To address this
problem, we introduce a deformable feature extractor (DFE) to extract more effic
ient features, and design a dual correlation attention mechanism (DCA) to deal w
ith its inherent local connectivity. Successfully, our proposed approach extents
the potential of RN to be position-aware of semantic objects by introducing onl
y a small number of parameters. We evaluate our approach on two major benchmark
datasets, i.e., Omniglot and Mini-Imagenet, and on both of the datasets our appr
oach achieves state-of-the-art performance. It's worth noting that our 5-way 1-s
hot result on Omniglot even outperforms the previous 5-way 5-shot results.
****

Multi-Adversarial Faster-RCNN for Unrestricted Object Detection
Zhenwei He, Lei Zhang; Proceedings of the IEEE/CVF International Conference on
Computer Vision (ICCV), 2019, pp. 6668-6677
Conventional object detection methods essentially suppose that the training and
testing data are collected from a restricted target domain with expensive labeli
ng cost. For alleviating the problem of domain dependency and cumbersome labelin
g, this paper proposes to detect objects in unrestricted environment by leveragi
ng domain knowledge trained from an auxiliary source domain with sufficient labe
ls. Specifically, we propose a multi-adversarial Faster-RCNN (MAF) framework for
unrestricted object detection, which inherently addresses domain disparity mini
mization for domain adaptation in feature representation. The paper merits are i
n three-fold: 1) With the idea that object detectors often becomes domain incomp
atible when image distribution resulted domain disparity appears, we propose a h
ierarchical domain feature alignment module, in which multiple adversarial domai
n classifier submodules for layer-wise domain feature confusion are designed; 2)
An information invariant scale reduction module (SRM) for hierarchical feature
map resizing is proposed for promoting the training efficiency of adversarial do
main adaptation; 3) In order to improve the domain adaptability, the aggregated
proposal features with detection results are feed into a proposed weighted gradi
ent reversal layer (WGRL) for characterizing hard confused domain samples. We ev
aluate our MAF on unrestricted tasks including Cityscapes, KITTI, Sim10k, etc. a
nd the experiments show the state-of-the-art performance over the existing detec
tors.
****

Object Guided External Memory Network for Video Object Detection
Hanming Deng, Yang Hua, Tao Song, Zongpu Zhang, Zhengui Xue, Ruhui Ma, Nei
l Robertson, Haibing Guan; Proceedings of the IEEE/CVF International Conference
on Computer Vision (ICCV), 2019, pp. 6678-6687
Video object detection is more challenging than image object detection because o
f the deteriorated frame quality. To enhance the feature representation, state-o
f-the-art methods propagate temporal information into the deteriorated frame by
aligning and aggregating entire feature maps from multiple nearby frames. Howeve
r, restricted by feature map's low storage-efficiency and vulnerable content-add
ress allocation, long-term temporal information is not fully stressed by these m
ethods. In this work, we propose the first object guided external memory network
for online video object detection. Storage-efficiency is handled by object guid
ed hard-attention to selectively store valuable features, and long-term informat
ion is protected when stored in an addressable external data matrix. A set of re
ad/write operations are designed to accurately propagate/allocate and delete mul
ti-level memory feature under object guidance. We evaluate our method on the Ima
geNet VID dataset and achieve state-of-the-art performance as well as good speed
-accuracy tradeoff. Furthermore, by visualizing the external memory, we show the
detailed object-level reasoning process across frames.
****

An Empirical Study of Spatial Attention Mechanisms in Deep Networks
Xizhou Zhu, Dazhi Cheng, Zheng Zhang, Stephen Lin, Jifeng Dai; Proceedings o
f the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 668
8-6697
Attention mechanisms have become a popular component in deep neural networks, ye

t there has been little examination of how different influencing factors and met
hods for computing attention from these factors affect performance. Toward a bet
ter general understanding of attention mechanisms, we present an empirical study
 that ablates various spatial attention elements within a generalized attention
formulation, encompassing the dominant Transformer attention as well as the prev
alent deformable convolution and dynamic convolution modules. Conducted on a var
iety of applications, the study yields significant findings about spatial attent
ion in deep networks, some of which run counter to conventional understanding. F
or example, we find that the query and key content comparison in Transformer att
ention is negligible for self-attention, but vital for encoder-decoder attention
. A proper combination of deformable convolution with key content only saliency
achieves the best accuracy-efficiency tradeoff in self-attention. Our results su
ggest that there exists much room for improvement in the design of attention mec
hanisms.
****

Attribute Attention for Semantic Disambiguation in Zero-Shot Learning
Yang Liu, Jishun Guo, Deng Cai, Xiaofei He; Proceedings of the IEEE/CVF Inter
national Conference on Computer Vision (ICCV), 2019, pp. 6698-6707
Zero-shot learning (ZSL) aims to accurately recognize unseen objects by learning
 mapping matrices that bridge the gap between visual information and semantic at
tributes. Previous works implicitly treat attributes equally in compatibility sc
ore while ignoring that they have different importance for discrimination, which
 leads to severe semantic ambiguity. Considering both low-level visual informati
on and global class-level features that relate to this ambiguity, we propose a p
ractical Latent Feature Guided Attribute Attention (LFGAA) framework to perform
object-based attribute attention for semantic disambiguation. By distracting sem
antic activation in dimensions that cause ambiguity, our method outperforms exis
ting state-of-the-art methods on AwA2, CUB and SUN datasets in both inductive an
d transductive settings.
****

CIIDefence: Defeating Adversarial Attacks by Fusing Class-Specific Image Inpaint
ing and Image Denoising
Puneet Gupta, Esa Rahtu; Proceedings of the IEEE/CVF International Conference o
n Computer Vision (ICCV), 2019, pp. 6708-6717
This paper presents a novel approach for protecting deep neural networks from ad
versarial attacks, i.e., methods that add well-crafted imperceptible modificatio
ns to the original inputs such that they are incorrectly classified with high co
nfidence. The proposed defence mechanism is inspired by the recent works mitigat
ing the adversarial disturbances by the means of image reconstruction and denois
ing. However, unlike the previous works, we apply the reconstruction only for sm
all and carefully selected image areas that are most influential to the current
classification outcome. The selection process is guided by the class activation
map responses obtained for multiple top-ranking class labels. The same regions a
re also the most prominent for the adversarial perturbations and hence most impo
rtant to purify. The resulting inpainting task is substantially more tractable t
han the full image reconstruction, while still being able to prevent the adversa
rial attacks. Furthermore, we combine the selective image inpainting with wavele
t based image denoising to produce a non differentiable layer that prevents atta
cker from using gradient backpropagation. Moreover, the proposed nonlinearity ca
nnot be easily approximated with simple differentiable alternative as demonstrat
ed in the experiments with Backward Pass Differentiable Approximation (BPDA) att
ack. Finally, we experimentally show that the proposed Class-specific Image Inpa
inting Defence (CIIDefence) is able to withstand several powerful adversarial at
tacks including the BPDA. The obtained results are consistently better compared
to the other recent defence approaches.
****

ThunderNet: Towards Real-Time Generic Object Detection on Mobile Devices
Zheng Qin, Zeming Li, Zhaoning Zhang, Yiping Bao, Gang Yu, Yuxing Peng, Ji
an Sun; Proceedings of the IEEE/CVF International Conference on Computer Vision
(ICCV), 2019, pp. 6718-6727

Real-time generic object detection on mobile platforms is a crucial but challeng ing computer vision task. Prior lightweight CNN-based detectors are inclined to use one-stage pipeline. In this paper, we investigate the effectiveness of two-s tage detectors in real-time generic detection and propose a lightweight two-stag e detector named ThunderNet. In the backbone part, we analyze the drawbacks in p revious lightweight backbones and present a lightweight backbone designed for ob ject detection. In the detection part, we exploit an extremely efficient RPN and detection head design. To generate more discriminative feature representation, we design two efficient architecture blocks, Context Enhancement Module and Spat ial Attention Module. At last, we investigate the balance between the input reso lution, the backbone, and the detection head. Benefit from the highly efficient backbone and detection part design, ThunderNet surpasses previous lightweight on e-stage detectors with only 40% of the computational cost on PASCAL VOC and COCO benchmarks. Without bells and whistles, ThunderNet runs at 24.1 fps on an ARM-b ased device with 19.2 AP on COCO. To the best of our knowledge, this is the firs t real-time detector reported on ARM platforms. Code will be released for paper reproduction.

****

Dual Student: Breaking the Limits of the Teacher in Semi-Supervised Learning

Zhanghan Ke, Daoye Wang, Qiong Yan, Jimmy Ren, Rynson W.H. Lau; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 67 28-6736

Recently, consistency-based methods have achieved state-of-the-art results in se mi-supervised learning (SSL). These methods always involve two roles, an explici t or implicit teacher model and a student model, and penalize predictions under different perturbations by a consistency constraint. However, the weights of the se two roles are tightly coupled since the teacher is essentially an exponential moving average (EMA) of the student. In this work, we show that the coupled EMA teacher causes a performance bottleneck. To address this problem, we introduce Dual Student, which replaces the teacher with another student. We also define a novel concept, stable sample, following which a stabilization constraint is desi gned for our structure to be trainable. Further, we discuss two variants of our method, which produce even higher performance. Extensive experiments show that o ur method improves the classification performance significantly on several main SSL benchmarks. Specifically, it reduces the error rate of the 13-layer CNN from 16.84% to 12.39% on CIFAR-10 with 1k labels and from 34.10% to 31.56% on CIFAR-100 with 10k labels. In addition, our method also achieves a clear improvement i n domain adaptation.

****

MVP Matching: A Maximum-Value Perfect Matching for Mining Hard Samples, With App lication to Person Re-Identification

Han Sun, Zhiyuan Chen, Shiyang Yan, Lin Xu; Proceedings of the IEEE/CVF Inter national Conference on Computer Vision (ICCV), 2019, pp. 6737-6747

How to correctly stress hard samples in metric learning is critical for visual r ecognition tasks, especially in challenging person re-ID applications. Pedestria ns across cameras with significant appearance variations are easily confused, wh ich could bias the learned metric and slow down the convergence rate. In this pa per, we propose a novel weighted complete bipartite graph based maximum-value pe rfect (MVP) matching for mining the hard samples from a batch of samples. It can emphasize the hard positive and negative sample pairs respectively, and thus re lieve adverse optimization and sample imbalance problems. We then develop a new batch-wise MVP matching based loss objective and combine it in an end-to-end dee p metric learning manner. It leads to significant improvements in both convergen ce rate and recognition performance. Extensive empirical results on five person re-ID benchmark datasets, i.e., Market-1501, CUHK03-Detected, CUHK03-Labeled, Du ke-MTMC, and MSMT17, demonstrate the superiority of the proposed method. It can accelerate the convergence rate significantly while achieving state-of-the-art p erformance. The source code of our method is available at https://github.com/IAA I-CVResearchGroup/MVP-metric.

****

Adaptive Context Network for Scene Parsing
Jun Fu, Jing Liu, Yuhang Wang, Yong Li, Yongjun Bao, Jinhui Tang, Hanqing Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6748-6757
Recent works attempt to improve scene parsing performance by exploring different levels of contexts, and typically train a well-designed convolutional network to exploit useful contexts across all pixels equally. However, in this paper, we find that the context demands are varying from different pixels or regions in each image. Based on this observation, we propose an Adaptive Context Network (ACNet) to capture the pixel-aware contexts by a competitive fusion of global context and local context according to different per-pixel demands. Specifically, when given a pixel, the global context demand is measured by the similarity between the global feature and its local feature, whose reverse value can also be used to measure the local context demand. We model the two demanding measurements by the proposed global context module and local context module, respectively, to generate their adaptive contextual features. Furthermore, we import multiple such modules to build several adaptive context blocks in different levels of network to obtain a coarse-to-fine result. Finally, comprehensive experimental evaluations demonstrate the effectiveness of the proposed ACNet, and new state-of-the-arts performances are achieved on all four public datasets, i.e. Cityscapes, ADE20K, PASCAL Context, and COCO Stuff.
****

Constructing Self-Motivated Pyramid Curriculums for Cross-Domain Semantic Segmentation: A Non-Adversarial Approach
Qing Lian, Fengmao Lv, Lixin Duan, Boqing Gong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6758-6767
We propose a new approach, called self-motivated pyramid curriculum domain adaptation (PyCDA), to facilitate the adaptation of semantic segmentation neural networks from synthetic source domains to real target domains. Our approach draws on an insight connecting two existing works: curriculum domain adaptation and self-training. Inspired by the former, PyCDA constructs a pyramid curriculum which contains various properties about the target domain. Those properties are mainly about the desired label distributions over the target domain images, image regions, and pixels. By enforcing the segmentation neural network to observe those properties, we can improve the network's generalization capability to the target domain. Motivated by the self-training, we infer this pyramid of properties by resorting to the semantic segmentation network itself. Unlike prior work, we do not need to maintain any additional models (e.g., logistic regression or discriminator networks) or to solve minmax problems which are often difficult to optimize. We report state-of-the-art results for the adaptation from both GTAV and SYNTHIA to Cityscapes, two popular settings in unsupervised domain adaptation for semantic segmentation.
****

SparseMask: Differentiable Connectivity Learning for Dense Image Prediction
Huikai Wu, Junge Zhang, Kaiqi Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6768-6777
In this paper, we aim at automatically searching an efficient network architecture for dense image prediction. Particularly, we follow the encoder-decoder style and focus on designing a connectivity structure for the decoder. To achieve that, we design a densely connected network with learnable connections, named Fully Dense Network, which contains a large set of possible final connectivity structures. We then employ gradient descent to search the optimal connectivity from the dense connections. The search process is guided by a novel loss function, which pushes the weight of each connection to be binary and the connections to be sparse. The discovered connectivity achieves competitive results on two segmentation datasets, while runs more than three times faster and requires less than half parameters compared to the state-of-the-art methods. An extensive experiment shows that the discovered connectivity is compatible with various backbones and generalizes well to other dense image prediction tasks.
****

Significance-Aware Information Bottleneck for Domain Adaptive Semantic Segmentation
Yawei Luo, Ping Liu, Tao Guan, Junqing Yu, Yi Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6778-6787
For unsupervised domain adaptation problems, the strategy of aligning the two domains in latent feature space through adversarial learning has achieved much progress in image classification, but usually fails in semantic segmentation tasks in which the latent representations are overcomplex. In this work, we equip the adversarial network with a "significance-aware information bottleneck (SIB)", to address the above problem. The new network structure, called SIBAN, enables a significance-aware feature purification before the adversarial adaptation, which eases the feature alignment and stabilizes the adversarial training course. In two domain adaptation tasks, i.e., GTA5 -> Cityscapes and SYNTHIA -> Cityscapes, we validate that the proposed method can yield leading results compared with other feature-space alternatives. Moreover, SIBAN can even match the state-of-the-art output-space methods in segmentation accuracy, while the latter are often considered to be better choices for domain adaptive segmentation task.
****

Relational Attention Network for Crowd Counting
Anran Zhang, Jiayi Shen, Zehao Xiao, Fan Zhu, Xiantong Zhen, Xianbin Cao, Ling Shao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6788-6797
Crowd counting is receiving rapidly growing research interests due to its potential application value in numerous real-world scenarios. However, due to various challenges such as occlusion, insufficient resolution and dynamic backgrounds, crowd counting remains an unsolved problem in computer vision. Density estimation is a popular strategy for crowd counting, where conventional density estimation methods perform pixel-wise regression without explicitly accounting the interdependence of pixels. As a result, independent pixel-wise predictions can be noisy and inconsistent. In order to address such an issue, we propose a Relational Attention Network (RANet) with a self-attention mechanism for capturing interdependence of pixels. The RANet enhances the self-attention mechanism by accounting both short-range and long-range interdependence of pixels, where we respectively denote these implementations as local self-attention (LSA) and global self-attention (GSA). We further introduce a relation module to fuse LSA and GSA to achieve more informative aggregated feature representations. We conduct extensive experiments on four public datasets, including ShanghaiTech A, ShanghaiTech B, UCF-CC-50 and UCF-QNRF. Experimental results on all datasets suggest RANet consistently reduces estimation errors and surpasses the state-of-the-art approaches by large margins.
****

ACFNet: Attentional Class Feature Network for Semantic Segmentation
Fan Zhang, Yanqin Chen, Zhihang Li, Zhibin Hong, Jingtuo Liu, Feifei Ma, Junyu Han, Errui Ding; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6798-6807
Recent works have made great progress in semantic segmentation by exploiting richer context, most of which are designed from a spatial perspective. In contrast to previous works, we present the concept of class center which extracts the global context from a categorical perspective. This class-level context describes the overall representation of each class in an image. We further propose a novel module, named Attentional Class Feature (ACF) module, to calculate and adaptively combine different class centers according to each pixel. Based on the ACF module, we introduce a coarse-to-fine segmentation network, called Attentional Class Feature Network (ACFNet), which can be composed of an ACF module and any off-the-shell segmentation network (base network). In this paper, we use two types of base networks to evaluate the effectiveness of ACFNet. We achieve new state-of-the-art performance of 81.85% mIoU on Cityscapes dataset with only finely annotated data used for training.
****

Frame-to-Frame Aggregation of Active Regions in Web Videos for Weakly Supervised

Semantic Segmentation
Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, Sungroh Yoon; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6808-6818
When a deep neural network is trained on data with only image-level labeling, the regions activated in each image tend to identify only a small region of the target object. We propose a method of using videos automatically harvested from the web to identify a larger region of the target object by using temporal information, which is not present in the static image. The temporal variations in a video allow different regions of the target object to be activated. We obtain an activated region in each frame of a video, and then aggregate the regions from successive frames into a single image, using a warping technique based on optical flow. The resulting localization maps cover more of the target object, and can then be used as proxy ground-truth to train a segmentation network. This simple approach outperforms existing methods under the same level of supervision, and even approaches relying on extra annotations. Based on VGG-16 and ResNet 101 backbones, our method achieves the mIoU of 65.0 and 67.4, respectively, on PASCAL VOC 2012 test images, which represents a new state-of-the-art.
****
Boundary-Aware Feature Propagation for Scene Segmentation
Henghui Ding, Xudong Jiang, Ai Qun Liu, Nadia Magnenat Thalmann, Gang Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6819-6829
In this work, we address the challenging issue of scene segmentation. To increase the feature similarity of the same object while keeping the feature discrimination of different objects, we explore to propagate information throughout the image under the control of objects' boundaries. To this end, we first propose to learn the boundary as an additional semantic class to enable the network to be aware of the boundary layout. Then, we propose unidirectional acyclic graphs (UAGs) to model the function of undirected cyclic graphs (UCGs), which structurize the image via building graphic pixel-by-pixel connections, in an efficient and effective way. Furthermore, we propose a boundary-aware feature propagation (BFP) module to harvest and propagate the local features within their regions isolated by the learned boundaries in the UAG-structured image. The proposed BFP is capable of splitting the feature propagation into a set of semantic groups via building strong connections among the same segment region but weak connections between different segment regions. Without bells and whistles, our approach achieves new state-of-the-art segmentation performance on three challenging semantic segmentation datasets, i.e., PASCAL-Context, CamVid, and Cityscapes.
****
Self-Ensembling With GAN-Based Data Augmentation for Domain Adaptation in Semantic Segmentation
Jaehoon Choi, Taekyung Kim, Changick Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6830-6840
Deep learning-based semantic segmentation methods have an intrinsic limitation that training a model requires a large amount of data with pixel-level annotations. To address this challenging issue, many researchers give attention to unsupervised domain adaptation for semantic segmentation. Unsupervised domain adaptation seeks to adapt the model trained on the source domain to the target domain. In this paper, we introduce a self-ensembling technique, one of the successful methods for domain adaptation in classification. However, applying self-ensembling to semantic segmentation is very difficult because heavily-tuned manual data augmentation used in self-ensembling is not useful to reduce the large domain gap in the semantic segmentation. To overcome this limitation, we propose a novel framework consisting of two components, which are complementary to each other. First, we present a data augmentation method based on Generative Adversarial Networks (GANs), which is computationally efficient and effective to facilitate domain alignment. Given those augmented images, we apply self-ensembling to enhance the performance of the segmentation network on the target domain. The proposed method outperforms state-of-the-art semantic segmentation methods on unsupervised do

main adaptation benchmarks.
****

Explaining the Ambiguity of Object Detection and 6D Pose From Visual Data
Fabian Manhardt, Diego Martin Arroyo, Christian Rupprecht, Benjamin Busam, T olga Birdal, Nassir Navab, Federico Tombari; Proceedings of the IEEE/CVF Inter national Conference on Computer Vision (ICCV), 2019, pp. 6841-6850
3D object detection and pose estimation from a single image are two inherently a mbiguous problems. Oftentimes, objects appear similar from different viewpoints due to shape symmetries, occlusion and repetitive textures. This ambiguity in bo th detection and pose estimation means that an object instance can be perfectly described by several different poses and even classes. In this work we propose t o explicitly deal with these ambiguities. For each object instance we predict mu ltiple 6D pose outcomes to estimate the specific pose distribution generated by symmetries and repetitive textures. The distribution collapses to a single outco me when the visual appearance uniquely identifies just one valid pose. We show t he benefits of our approach which provides not only a better explanation for pos e ambiguity, but also a higher accuracy in terms of pose estimation.
****

Accurate Monocular 3D Object Detection via Color-Embedded 3D Reconstruction for Autonomous Driving
Xinzhu Ma, Zhihui Wang, Haojie Li, Pengbo Zhang, Wanli Ouyang, Xin Fan; Pro ceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 201 9, pp. 6851-6860
In this paper, we propose a monocular 3D object detection framework in the domai n of autonomous driving. Unlike previous image-based methods which focus on RGB feature extracted from 2D images, our method solves this problem in the reconstr ucted 3D space in order to exploit 3D contexts explicitly. To this end, we first leverage a stand-alone module to transform the input data from 2D image plane t o 3D point clouds space for a better input representation, then we perform the 3 D detection using PointNet backbone net to obtain objects' 3D locations, dimensi ons and orientations. To enhance the discriminative capability of point clouds, we propose a multi-modal feature fusion module to embed the complementary RGB cu e into the generated point clouds representation. We argue that it is more effec tive to infer the 3D bounding boxes from the generated 3D scene space (i.e., X,Y , Z space) compared to the image plane (i.e., R,G,B image plane). Evaluation on the challenging KITTI dataset shows that our approach boosts the performance of state-of-the-art monocular approach by a large margin.
****

MonoLoco: Monocular 3D Pedestrian Localization and Uncertainty Estimation
Lorenzo Bertoni, Sven Kreiss, Alexandre Alahi; Proceedings of the IEEE/CVF Int ernational Conference on Computer Vision (ICCV), 2019, pp. 6861-6871
We tackle the fundamentally ill-posed problem of 3D human localization from mono cular RGB images. Driven by the limitation of neural networks outputting point e stimates, we address the ambiguity in the task by predicting confidence interval s through a loss function based on the Laplace distribution. Our architecture is a light-weight feed-forward neural network that predicts 3D locations and corre sponding confidence intervals given 2D human poses. The design is particularly w ell suited for small training data, cross-dataset generalization, and real-time applications. Our experiments show that we (i) outperform state-of-the-art resul ts on KITTI and nuScenes datasets, (ii) even outperform a stereo-based method fo r far-away pedestrians, and (iii) estimate meaningful confidence intervals. We f urther share insights on our model of uncertainty in cases of limited observatio ns and out-of-distribution samples.
****

Unsupervised High-Resolution Depth Learning From Videos With Dual Networks
Junsheng Zhou, Yuwang Wang, Kaihuai Qin, Wenjun Zeng; Proceedings of the IEEE /CVF International Conference on Computer Vision (ICCV), 2019, pp. 6872-6881
Unsupervised depth learning takes the appearance difference between a target vie w and a view synthesized from its adjacent frame as supervisory signal. Since th e supervisory signal only comes from images themselves, the resolution of traini

ng data significantly impacts the performance. High-resolution images contain mo re fine-grained details and provide more accurate supervisory signal. However, d ue to the limitation of memory and computation power, the original images are ty pically down-sampled during training, which suffers heavy loss of details and di sparity accuracy. In order to fully explore the information contained in high-re solution data, we propose a simple yet effective dual networks architecture, whi ch can directly take high-resolution images as input and generate high-resolutio n and high-accuracy depth map efficiently. We also propose a Self-assembled Atte ntion (SA-Attention) module to handle low-texture region. The evaluation on the benchmark KITTI and Make3D datasets demonstrates that our method achieves state-of-the-art results in the monocular depth estimation task.
****

Bayesian Graph Convolution LSTM for Skeleton Based Action Recognition
Rui Zhao, Kang Wang, Hui Su, Qiang Ji; Proceedings of the IEEE/CVF Internatio nal Conference on Computer Vision (ICCV), 2019, pp. 6882-6892
We propose a framework for recognizing human actions from skeleton data by model ing the underlying dynamic process that generates the motion pattern. We capture three major factors that contribute to the complexity of the motion pattern inc luding spatial dependencies among body joints, temporal dependencies of body pos es, and variation among subjects in action execution. We utilize graph convoluti on to extract structure-aware feature representation from pose data by exploitin g the skeleton anatomy. Long short-term memory (LSTM) network is then used to ca pture the temporal dynamics of the data. Finally, the whole model is extended un der the Bayesian framework to a probabilistic model in order to better capture t he stochasticity and variation in the data. An adversarial prior is developed to regularize the model parameters to improve the generalization of the model. A B ayesian inference problem is formulated to solve the classification task. We dem onstrate the benefit of this framework in several benchmark datasets with recogn ition under various generalization conditions.
****

DeCaFA: Deep Convolutional Cascade for Face Alignment in the Wild
Arnaud Dapogny, Kevin Bailly, Matthieu Cord; Proceedings of the IEEE/CVF Inter national Conference on Computer Vision (ICCV), 2019, pp. 6893-6901
Face Alignment is an active computer vision domain, that consists in localizing a number of facial landmarks that vary across datasets. State-of-the-art face al ignment methods either consist in end-to-end regression, or in refining the shap e in a cascaded manner, starting from an initial guess. In this paper, we introd uce an end-to-end deep convolutional cascade (DeCaFA) architecture for face alig nment. Face Alignment is an active computer vision domain, that consists in loca lizing a number of facial landmarks that vary across datasets. State-of-the-art face alignment methods either consist in end-to-end regression, or in refining t he shape in a cascaded manner, starting from an initial guess. In this paper, we introduce DeCaFA, an end-to-end deep convolutional cascade architecture for fac e alignment. DeCaFA uses fully-convolutional stages to keep full spatial resolut ion throughout the cascade. Between each cascade stage, DeCaFA uses multiple cha ined transfer layers with spatial softmax to produce landmark-wise attention map s for each of several landmark alignment tasks. Weighted intermediate supervisio n, as well as efficient feature fusion between the stages allow to learn to prog ressively refine the attention maps in an end-to-end manner. We show experimenta lly that DeCaFA significantly outperforms existing approaches on 300W, CelebA an d WFLW databases. In addition, we show that DeCaFA can learn fine alignment with reasonable accuracy from very few images using coarsely annotated data.
****

Probabilistic Face Embeddings
Yichun Shi, Anil K. Jain; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6902-6911
Embedding methods have achieved success in face recognition by comparing facial features in a latent semantic space. However, in a fully unconstrained face sett ing, the facial features learned by the embedding model could be ambiguous or ma y not even be present in the input face, leading to noisy representations. We pr

opose Probabilistic Face Embeddings (PFEs), which represent each face image as a Gaussian distribution in the latent space. The mean of the distribution estimates the most likely feature values while the variance shows the uncertainty in the feature values. Probabilistic solutions can then be naturally derived for matching and fusing PFEs using the uncertainty information. Empirical evaluation on different baseline models, training datasets and benchmarks show that the proposed method can improve the face recognition performance of deterministic embeddings by converting them into PFEs. The uncertainties estimated by PFEs also serve as good indicators of the potential matching accuracy, which are important for a risk-controlled recognition system.
****

Gaze360: Physically Unconstrained Gaze Estimation in the Wild
Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, Antonio Torralba; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6912-6921
Understanding where people are looking is an informative social cue. In this work, we present Gaze360, a large-scale remote gaze-tracking dataset and method for robust 3D gaze estimation in unconstrained images. Our dataset consists of 238 subjects in indoor and outdoor environments with labelled 3D gaze across a wide range of head poses and distances. It is the largest publicly available dataset of its kind by both subject and variety, made possible by a simple and efficient collection method. Our proposed 3D gaze model extends existing models to include temporal information and to directly output an estimate of gaze uncertainty. We demonstrate the benefits of our model via an ablation study, and show its generalization performance via a cross-dataset evaluation against other recent gaze benchmark datasets. We furthermore propose a simple self-supervised approach to improve cross-dataset domain adaptation. Finally, we demonstrate an application of our model for estimating customer attention in a supermarket setting. Our dataset and models will be made available at http://gaze360.csail.mit.edu.
****

Unsupervised Person Re-Identification by Camera-Aware Similarity Consistency Learning
Ancong Wu, Wei-Shi Zheng, Jian-Huang Lai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6922-6931
For matching pedestrians across disjoint camera views in surveillance, person re-identification (Re-ID) has made great progress in supervised learning. However, it is infeasible to label data in a number of new scenes when extending a Re-ID system. Thus, studying unsupervised learning for Re-ID is important for saving labelling cost. Yet, cross-camera scene variation is a key challenge for unsupervised Re-ID, such as illumination, background and viewpoint variations, which cause domain shift in the feature space and result in inconsistent pairwise similarity distributions that degrade matching performance. To alleviate the effect of cross-camera scene variation, we propose a Camera-Aware Similarity Consistency Loss to learn consistent pairwise similarity distributions for intra-camera matching and cross-camera matching. To avoid learning ineffective knowledge in consistency learning, we preserve the prior common knowledge of intra-camera matching in the pretrained model as reliable guiding information, which does not suffer from cross-camera scene variation as cross-camera matching. To learn similarity consistency more effectively, we further develop a coarse-to-fine consistency learning scheme to learn consistency globally and locally in two steps. Experiments show that our method outperformed the state-of-the-art unsupervised Re-ID methods.
****

Photo-Realistic Monocular Gaze Redirection Using Generative Adversarial Networks
Zhe He, Adrian Spurr, Xucong Zhang, Otmar Hilliges; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6932-6941
Gaze redirection is the task of changing the gaze to a desired direction for a given monocular eye patch image. Many applications such as videoconferencing, films, games, and generation of training data for gaze estimation require redirecting the gaze, without distorting the appearance of the area surrounding the eye a

nd while producing photo-realistic images. Existing methods lack the ability to generate perceptually plausible images. In this work, we present a novel method to alleviate this problem by leveraging generative adversarial training to synth esize an eye image conditioned on a target gaze direction. Our method ensures pe rceptual similarity and consistency of synthesized images to the real images. Fu rthermore, a gaze estimation loss is used to control the gaze direction accurate ly. To attain high-quality images, we incorporate perceptual and cycle consisten cy losses into our architecture. In extensive evaluations we show that the propo sed method outperforms state-of-the-art approaches in terms of both image qualit y and redirection precision. Finally, we show that generated images can bring si gnificant improvement for the gaze estimation task if used to augment real train ing data.

****

Dynamic Kernel Distillation for Efficient Pose Estimation in Videos

Xuecheng Nie, Yuncheng Li, Linjie Luo, Ning Zhang, Jiashi Feng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6942-6950

Existing video-based human pose estimation methods extensively apply large netwo rks onto every frame in the video to localize body joints, which suffer high com putational cost and hardly meet the low-latency requirement in realistic applica tions. To address this issue, we propose a novel Dynamic Kernel Distillation (DK D) model to facilitate small networks for estimating human poses in videos, thus significantly lifting the efficiency. In particular, DKD introduces a light-wei ght distillator to online distill pose kernels via leveraging temporal cues from the previous frame in a one-shot feed-forward manner. Then, DKD simplifies body joint localization into a matching procedure between the pose kernels and the c urrent frame, which can be efficiently computed via simple convolution. In this way, DKD fast transfers pose knowledge from one frame to provide compact guidanc e for body joint localization in the following frame, which enables utilization of small networks in video-based pose estimation. To facilitate the training pro cess, DKD exploits a temporally adversarial training strategy that introduces a temporal discriminator to help generate temporally coherent pose kernels and pos e estimation results within a long range. Experiments on Penn Action and Sub-JHM DB benchmarks demonstrate outperforming efficiency of DKD, specifically, 10x flo ps reduction and 2x speedup over previous best model, and its state-of-the-art a ccuracy.

****

Single-Stage Multi-Person Pose Machines

Xuecheng Nie, Jiashi Feng, Jianfeng Zhang, Shuicheng Yan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6951-6960

Multi-person pose estimation is a challenging problem. Existing methods are most ly two-stage based-one stage for proposal generation and the other for allocatin g poses to corresponding persons. However, such two-stage methods generally suff er low efficiency. In this work, we present the first single-stage model, Single -stage multi-person Pose Machine (SPM), to simplify the pipeline and lift the ef ficiency for multi-person pose estimation. To achieve this, we propose a novel S tructured Pose Representation (SPR) that unifies person instance and body joint position representations. Based on SPR, we develop the SPM model that can direct ly predict structured poses for multiple persons in a single stage, and thus off er a more compact pipeline and attractive efficiency advantage over two-stage me thods. In particular, SPR introduces the root joints to indicate different perso n instances and human body joint positions are encoded into their displacements w.r.t. the roots. To better predict long-range displacements for some joints, SP R is further extended to hierarchical representations. Based on SPR, SPM can eff iciently perform multi-person poses estimation by simultaneously predicting root joints (location of instances) and body joint displacements via CNNs. Moreover, to demonstrate the generality of SPM, we also apply it to multi-person 3D pose estimation. Comprehensive experiments on benchmarks MPII, extended PASCAL-Person -Part, MSCOCO and CMU Panoptic clearly demonstrate the state-of-the-art efficien cy of SPM for multi-person 2D/3D pose estimation, together with outstanding accu

racy.
****
SO-HandNet: Self-Organizing Network for 3D Hand Pose Estimation With Semi-Supervised Learning
Yujin Chen, Zhigang Tu, Liuhao Ge, Dejun Zhang, Ruizhi Chen, Junsong Yuan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6961-6970
3D hand pose estimation has made significant progress recently, where Convolutional Neural Networks (CNNs) play a critical role. However, most of the existing CNN-based hand pose estimation methods depend much on the training set, while labeling 3D hand pose on training data is laborious and time-consuming. Inspired by the point cloud autoencoder presented in self-organizing network (SO-Net), our proposed SO-HandNet aims at making use of the unannotated data to obtain accurate 3D hand pose estimation in a semi-supervised manner. We exploit hand feature encoder (HFE) to extract multi-level features from hand point cloud and then fuse them to regress 3D hand pose by a hand pose estimator (HPE). We design a hand feature decoder (HFD) to recover the input point cloud from the encoded feature. Since the HFE and the HFD can be trained without 3D hand pose annotation, the proposed method is able to make the best of unannotated data during the training phase. Experiments on four challenging benchmark datasets validate that our proposed SO-HandNet can achieve superior performance for 3D hand pose estimation via semi-supervised learning.
****
Adaptive Wing Loss for Robust Face Alignment via Heatmap Regression
Xinyao Wang, Liefeng Bo, Li Fuxin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6971-6981
Heatmap regression with a deep network has become one of the mainstream approaches to localize facial landmarks. However, the loss function for heatmap regression is rarely studied. In this paper, we analyze the ideal loss function properties for heatmap regression in face alignment problems. Then we propose a novel loss function, named Adaptive Wing loss, that is able to adapt its shape to different types of ground truth heatmap pixels. This adaptability penalizes loss more on foreground pixels while less on background pixels. To address the imbalance between foreground and background pixels, we also propose Weighted Loss Map, which assigns high weights on foreground and difficult background pixels to help training process focus more on pixels that are crucial to landmark localization. To further improve face alignment accuracy, we introduce boundary prediction and CoordConv with boundary coordinates. Extensive experiments on different benchmarks, including COFW, 300W and WFLW, show our approach outperforms the state-of-the-art by a significant margin on various evaluation metrics. Besides, the Adaptive Wing loss also helps other heatmap regression tasks.
****
Single-Network Whole-Body Pose Estimation
Gines Hidalgo, Yaadhav Raaj, Haroon Idrees, Donglai Xiang, Hanbyul Joo, Tomas Simon, Yaser Sheikh; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6982-6991
We present the first single-network approach for 2D whole-body pose estimation, which entails simultaneous localization of body, face, hands, and feet keypoints. Due to the bottom-up formulation, our method maintains constant real-time performance regardless of the number of people in the image. The network is trained in a single stage using multi-task learning, through an improved architecture which can handle scale differences between body/foot and face/hand keypoints. Our approach considerably improves upon OpenPose [??], the only work so far capable of whole-body pose estimation, both in terms of speed and global accuracy. Unlike OpenPose, our method does not need to run an additional network for each hand and face candidate, making it substantially faster for multi-person scenarios. This work directly results in a reduction of computational complexity for applications that require 2D whole-body information (e.g., VR/AR, re-targeting). In addition, it yields higher accuracy, especially for occluded, blurry, and low resolution faces and hands. For code, trained models, and validation benchmarks, visi

t our project page: https://github.com/CMU-Perceptual-Computing-Lab/openpose_train.
****

Face Alignment With Kernel Density Deep Neural Network
Lisha Chen, Hui Su, Qiang Ji; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6992-7002
Deep neural networks achieve good performance in many computer vision problems such as face alignment. However, when the testing image is challenging due to low resolution, occlusion or adversarial attacks, the accuracy of a deep neural network suffers greatly. Therefore, it is important to quantify the uncertainty in its predictions. A probabilistic neural network with Gaussian distribution over the target is typically used to quantify uncertainty for regression problems. However, in real-world problems especially computer vision tasks, the Gaussian assumption is too strong. To model more general distributions, such as multi-modal or asymmetric distributions, we propose to develop a kernel density deep neural network. Specifically, for face alignment, we adapt state-of-the-art hourglass neural network into a probabilistic neural network framework with landmark probability map as its output. The model is trained by maximizing the conditional log likelihood. To exploit the output probability map, we extend the model to multi-stage so that the logits map from the previous stage can feed into the next stage to progressively improve the landmark detection accuracy. Extensive experiments on benchmark datasets against state-of-the-art unconstrained deep learning method demonstrate that the proposed kernel density network achieves comparable or superior performance in terms of prediction accuracy. It further provides aleatoric uncertainty estimation in predictions.
****

Spatiotemporal Feature Residual Propagation for Action Prediction
He Zhao, Richard P. Wildes; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7003-7012
Recognizing actions from limited preliminary video observations has seen considerable recent progress. Typically, however, such progress has been had without explicitly modeling fine-grained motion evolution as a potentially valuable information source. In this study, we address this task by investigating how action patterns evolve over time in a spatial feature space. There are three key components to our system. First, we work with intermediate-layer ConvNet features, which allow for abstraction from raw data, while retaining spatial layout, which is sacrificed in approaches that rely on vectorized global representations. Second, instead of propagating features per se, we propagate their residuals across time, which allows for a compact representation that reduces redundancy while retaining essential information about evolution over time. Third, we employ a Kalman filter to combat error build-up and unify across prediction start times. Extensive experimental results on the JHMDB21, UCF101 and BIT datasets show that our approach leads to a new state-of-the-art in action prediction.
****

Identity From Here, Pose From There: Self-Supervised Disentanglement and Generation of Objects Using Unlabeled Videos
Fanyi Xiao, Haotian Liu, Yong Jae Lee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7013-7022
We propose a novel approach that disentangles the identity and pose of objects for image generation. Our model takes as input an ID image and a pose image, and generates an output image with the identity of the ID image and the pose of the pose image. Unlike most previous unsupervised work which rely on cyclic constraints, which can often be brittle, we instead propose to learn this in a self-supervised way. Specifically, we leverage unlabeled videos to automatically construct pseudo ground-truth targets to directly supervise our model. To enforce disentanglement, we propose a novel disentanglement loss, and to improve realism, we propose a pixel-verification loss in which the generated image's pixels must trace back to the ID input. We conduct extensive experiments on both synthetic and real images to demonstrate improved realism, diversity, and ID/pose disentanglement compared to existing methods.

****
Relation Distillation Networks for Video Object Detection
Jiajun Deng, Yingwei Pan, Ting Yao, Wengang Zhou, Houqiang Li, Tao Mei; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7023-7032
It has been well recognized that modeling object-to-object relations would be helpful for object detection. Nevertheless, the problem is not trivial especially when exploring the interactions between objects to boost video object detectors. The difficulty originates from the aspect that reliable object relations in a video should depend on not only the objects in the present frame but also all the supportive objects extracted over a long range span of the video. In this paper, we introduce a new design to capture the interactions across the objects in spatio-temporal context. Specifically, we present Relation Distillation Networks (RDN) --- a new architecture that novelly aggregates and propagates object relation to augment object features for detection. Technically, object proposals are first generated via Region Proposal Networks (RPN). RDN then, on one hand, models object relation via multi-stage reasoning, and on the other, progressively distills relation through refining supportive object proposals with high objectness scores in a cascaded manner. The learnt relation verifies the efficacy on both improving object detection in each frame and box linking across frames. Extensive experiments are conducted on ImageNet VID dataset, and superior results are reported when comparing to state-of-the-art methods. More remarkably, our RDN achieves 81.8% and 83.2% mAP with ResNet-101 and ResNeXt-101, respectively. When further equipped with linking and rescoring, we obtain to-date the best reported mAP of 83.8% and 84.7%.
****
Video Compression With Rate-Distortion Autoencoders
Amirhossein Habibian, Ties van Rozendaal, Jakub M. Tomczak, Taco S. Cohen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7033-7042
In this paper we present a a deep generative model for lossy video compression. We employ a model that consists of a 3D autoencoder with a discrete latent space and an autoregressive prior used for entropy coding. Both autoencoder and prior are trained jointly to minimize a rate-distortion loss, which is closely related to the ELBO used in variational autoencoders. Despite its simplicity, we find that our method outperforms the state-of-the-art learned video compression networks based on motion compensation or interpolation. We systematically evaluate various design choices, such as the use of frame-based or spatio-temporal autoencoders, and the type of autoregressive prior. In addition, we present three extensions of the basic method that demonstrate the benefits over classical approaches to compression. First, we introduce semantic compression, where the model is trained to allocate more bits to objects of interest. Second, we study adaptive compression, where the model is adapted to a domain with limited variability, e.g. videos taken from an autonomous car, to achieve superior compression on that domain. Finally, we introduce multimodal compression, where we demonstrate the effectiveness of our model in joint compression of multiple modalities captured by non-standard imaging sensors, such as quad cameras. We believe that this opens up novel video compression applications, which have not been feasible with classical codecs.
****
Non-Local ConvLSTM for Video Compression Artifact Reduction
Yi Xu, Longwen Gao, Kai Tian, Shuigeng Zhou, Huyang Sun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7043-7052
Video compression artifact reduction aims to recover high-quality videos from low-quality compressed videos. Most existing approaches use a single neighboring frame or a pair of neighboring frames (preceding and/or following the target frame) for this task. Furthermore, as frames of high quality overall may contain low-quality patches, and high-quality patches may exist in frames of low quality overall, current methods focusing on nearby peak-quality frames (PQFs) may miss high-quality details in low-quality frames. To remedy these shortcomings, in this

paper we propose a novel end-to-end deep neural network called non-local ConvLSTM (NL-ConvLSTM in short) that exploits multiple consecutive frames. An approximate non-local strategy is introduced in NL-ConvLSTM to capture global motion patterns and trace the spatiotemporal dependency in a video sequence. This approximate strategy makes the non-local module work in a fast and low space-cost way. Our method uses the preceding and following frames of the target frame to generate a residual, from which a higher quality frame is reconstructed. Experiments on two datasets show that NL-ConvLSTM outperforms the existing methods.
****

Self-Supervised Moving Vehicle Tracking With Stereo Sound
Chuang Gan, Hang Zhao, Peihao Chen, David Cox, Antonio Torralba; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7053-7062
Humans are able to localize objects in the environment using both visual and auditory cues, integrating information from multiple modalities into a common reference frame. We introduce a system that can leverage unlabeled audiovisual data to learn to localize objects (moving vehicles) in a visual reference frame, purely using stereo sound at inference time. Since it is labor-intensive to manually annotate the correspondences between audio and object bounding boxes, we achieve this goal by using the co-occurrence of visual and audio streams in unlabeled videos as a form of self-supervision, without resorting to the collection of ground truth annotations. In particular, we propose a framework that consists of a vision "teacher" network and a stereo-sound "student" network. During training, knowledge embodied in a well-established visual vehicle detection model is transferred to the audio domain using unlabeled videos as a bridge. At test time, the stereo-sound student network can work independently to perform object localization using just stereo audio and camera meta-data, without any visual input. Experimental results on a newly collected Auditory Vehicles Tracking dataset verify that our proposed approach outperforms several baseline approaches. We also demonstrate that our cross-modal auditory localization approach can assist in the visual localization of moving vehicles under poor lighting conditions.
****

Self-Supervised Learning With Geometric Constraints in Monocular Video: Connecting Flow, Depth, and Camera
Yuhua Chen, Cordelia Schmid, Cristian Sminchisescu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7063-7072
We present GLNet, a self-supervised framework for learning depth, optical flow, camera pose and intrinsic parameters from monocular video -- addressing the difficulty of acquiring realistic ground-truth for such tasks. We propose three contributions: 1) we design new loss functions that capture multiple geometric constraints (eg. epipolar geometry) as well as adaptive photometric loss that supports multiple moving objects, rigid and non-rigid, 2) we extend the model such that it predicts camera intrinsics, making it applicable to uncalibrated video, and 3) we propose several online refinement strategies that rely on the symmetry of our self-supervised loss in training and testing, in particular optimizing model parameters and/or the output of different tasks, leveraging their mutual interactions. The idea of jointly optimizing the system output, under all geometric and photometric constraints can be viewed as a dense generalization of classical bundle adjustment. We demonstrate the effectiveness of our method on KITTI and Cityscapes, where we outperform previous self-supervised approaches on multiple tasks. We also show good generalization for transfer learning.
****

Learning Temporal Action Proposals With Fewer Labels
Jingwei Ji, Kaidi Cao, Juan Carlos Niebles; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7073-7082
Temporal action proposals are a common module in action detection pipelines today. Most current methods for training action proposal modules rely on fully supervised approaches that require large amounts of annotated temporal action intervals in long video sequences. The large cost and effort in annotation that this entails motivate us to study the problem of training proposal modules with less su

pervision. In this work, we propose a semi-supervised learning algorithm specifically designed for training temporal action proposal networks. When only a small number of labels are available, our semi-supervised method generates significantly better proposals than the fully-supervised counterpart and other strong semi-supervised baselines. We validate our method on two challenging action detection video datasets, ActivityNet v1.3 and THUMOS14. We show that our semi-supervised approach consistently matches or outperforms the fully supervised state-of-the-art approaches.
****

TSM: Temporal Shift Module for Efficient Video Understanding
Ji Lin, Chuang Gan, Song Han; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7083-7093
The explosive growth in video streaming gives rise to challenges on performing video understanding at high accuracy and low computation cost. Conventional 2D CNNs are computationally cheap but cannot capture temporal relationships; 3D CNN based methods can achieve good performance but are computationally intensive, making it expensive to deploy. In this paper, we propose a generic and effective Temporal Shift Module (TSM) that enjoys both high efficiency and high performance. Specifically, it can achieve the performance of 3D CNN but maintain 2D CNN's complexity. TSM shifts part of the channels along the temporal dimension; thus facilitate information exchanged among neighboring frames. It can be inserted into 2D CNNs to achieve temporal modeling at zero computation and zero parameters. We also extended TSM to online setting, which enables real-time low-latency online video recognition and video object detection. TSM is accurate and efficient: it ranks the first place on the Something-Something leaderboard upon publication; on Jetson Nano and Galaxy Note8, it achieves a low latency of 13ms and 35ms for online video recognition. The code is available at: https://github. com/mit-han-lab/temporal-shift-module.
****

Graph Convolutional Networks for Temporal Action Localization
Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, Chuang Gan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7094-7103
Most state-of-the-art action localization systems process each action proposal individually, without explicitly exploiting their relations during learning. However, the relations between proposals actually play an important role in action localization, since a meaningful action always consists of multiple proposals in a video. In this paper, we propose to exploit the proposal-proposal relations using GraphConvolutional Networks (GCNs). First, we construct an action proposal graph, where each proposal is represented as a node and their relations between two proposals as an edge. Here, we use two types of relations, one for capturing the context information for each proposal and the other one for characterizing the correlations between distinct actions. Then we apply the GCNs over the graph to model the relations among different proposals and learn powerful representations for the action classification and localization. Experimental results show that our approach significantly outperforms the state-of-the-art on THUMOS14(49.1% versus 42.8%). Moreover, augmentation experiments on ActivityNet also verify the efficacy of modeling action proposal relationships.
****

Fast Object Detection in Compressed Video
Shiyao Wang, Hongchao Lu, Zhidong Deng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7104-7113
Object detection in videos has drawn increasing attention since it is more practical in real scenarios. Most of the deep learning methods use CNNs to process each decoded frame in a video stream individually. However, the free of charge yet valuable motion information already embedded in the video compression format is usually overlooked. In this paper, we propose a fast object detection method by taking advantage of this with a novel Motion aided Memory Network (MMNet). The MMNet has two major advantages: 1) It significantly accelerates the procedure of feature extraction for compressed videos. It only need to run a complete recogn

ition network for I-frames, i.e. a few reference frames in a video, and it produces the features for the following P frames (predictive frames) with a light weight memory network, which runs fast; 2) Unlike existing methods that establish an additional network to model motion of frames, we take full advantage of both motion vectors and residual errors that are freely available in video streams. To our best knowledge, the MMNet is the first work that investigates a deep convolutional detector on compressed videos. Our method is evaluated on the large-scale ImageNet VID dataset, and the results show that it is 3x times faster than single image detector R-FCN and 10x times faster than high-performance detector MANet at a minor accuracy loss.

****

Predicting 3D Human Dynamics From Video

Jason Y. Zhang, Panna Felsen, Angjoo Kanazawa, Jitendra Malik; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7114-7123

Given a video of a person in action, we can easily guess the 3D future motion of the person. In this work, we present perhaps the first approach for predicting a future 3D mesh model sequence of a person from past video input. We do this for periodic motions such as walking and also actions like bowling and squatting seen in sports or workout videos. While there has been a surge of future prediction problems in computer vision, most approaches predict 3D future from 3D past or 2D future from 2D past inputs. In this work, we focus on the problem of predicting 3D future motion from past image sequences, which has a plethora of practical applications in autonomous systems that must operate safely around people from visual inputs. Inspired by the success of autoregressive models in language modeling tasks, we learn an intermediate latent space on which we predict the future. This effectively facilitates autoregressive predictions when the input differs from the output domain. Our approach can be trained on video sequences obtained in-the-wild without 3D ground truth labels. The project website with videos can be found at https://jasonyzhang.com/phd.

****

Imitation Learning for Human Pose Prediction

Borui Wang, Ehsan Adeli, Hsu-kuang Chiu, De-An Huang, Juan Carlos Niebles; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7124-7133

Modeling and prediction of human motion dynamics has long been a challenging problem in computer vision, and most existing methods rely on the end-to-end supervised training of various architectures of recurrent neural networks. Inspired by the recent success of deep reinforcement learning methods, in this paper we propose a new reinforcement learning formulation for the problem of human pose prediction, and develop an imitation learning algorithm for predicting future poses under this formulation through a combination of behavioral cloning and generative adversarial imitation learning. Our experiments show that our proposed method outperforms all existing state-of-the-art baseline models by large margins on the task of human pose prediction in both short-term predictions and long-term predictions, while also enjoying huge advantage in training speed.

****

Human Motion Prediction via Spatio-Temporal Inpainting

Alejandro Hernandez, Jurgen Gall, Francesc Moreno-Noguer; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7134-7143

We propose a Generative Adversarial Network (GAN) to forecast 3D human motion given a sequence of past 3D skeleton poses. While recent GANs have shown promising results, they can only forecast plausible motion over relatively short periods of time (few hundred milliseconds) and typically ignore the absolute position of the skeleton w.r.t. the camera. Our scheme provides long term predictions (two seconds or more) for both the body pose and its absolute position. Our approach builds upon three main contributions. First, we represent the data using a spatio-temporal tensor of 3D skeleton coordinates which allows formulating the prediction problem as an inpainting one, for which GANs work particularly well. Secondly, we design an architecture to learn the joint distribution of body poses and

global motion, capable to hypothesize large chunks of the input 3D tensor with m
issing data. And finally, we argue that the L2 metric, considered so far by most
 approaches, fails to capture the actual distribution of long-term human motion.
 We propose two alternative metrics, based on the distribution of frequencies, t
hat are able to capture more realistic motion patterns. Extensive experiments de
monstrate our approach to significantly improve the state of the art, while also
 handling situations in which past observations are corrupted by occlusions, noi
se and missing frames.
****

Structured Prediction Helps 3D Human Motion Modelling
Emre Aksan, Manuel Kaufmann, Otmar Hilliges; Proceedings of the IEEE/CVF Inter
national Conference on Computer Vision (ICCV), 2019, pp. 7144-7153
Human motion prediction is a challenging and important task in many computer vis
ion application domains. Existing work only implicitly models the spatial struct
ure of the human skeleton. In this paper, we propose a novel approach that decom
poses the prediction into individual joints by means of a structured prediction
 layer that explicitly models the joint dependencies. This is implemented via a h
ierarchy of small-sized neural networks connected analogously to the kinematic c
hains in the human body as well as a joint-wise decomposition in the loss functi
on. The proposed layer is agnostic to the underlying network and can be used wit
h existing architectures for motion modelling. Prior work typically leverages th
e H3.6M dataset. We show that some state-of-the-art techniques do not perform we
ll when trained and tested on AMASS, a recently released dataset 14 times the si
ze of H3.6M. Our experiments indicate that the proposed layer increases the perf
ormance of motion forecasting irrespective of the base network, joint-angle repr
esentation, and prediction horizon. We furthermore show that the layer also impr
oves motion predictions qualitatively. We make code and models publicly availabl
e at https://ait.ethz.ch/projects/2019/spl.
****

Learning Shape Templates With Structured Implicit Functions
Kyle Genova, Forrester Cole, Daniel Vlasic, Aaron Sarna, William T. Freeman,
  Thomas Funkhouser; Proceedings of the IEEE/CVF International Conference on Com
puter Vision (ICCV), 2019, pp. 7154-7164
Template 3D shapes are useful for many tasks in graphics and vision, including f
itting observation data, analyzing shape collections, and transferring shape att
ributes. Because of the variety of geometry and topology of real-world shapes, p
revious methods generally use a library of hand-made templates. In this paper, w
e investigate learning a general shape template from data. To allow for widely v
arying geometry and topology, we choose an implicit surface representation based
 on composition of local shape elements. While long known to computer graphics,
this representation has not yet been explored in the context of machine learning
 for vision. We show that structured implicit functions are suitable for learnin
g and allow a network to smoothly and simultaneously fit multiple classes of sha
pes. The learned shape template supports applications such as shape exploration,
 correspondence, abstraction, interpolation, and semantic segmentation from an R
GB image.
****

CompenNet++: End-to-End Full Projector Compensation
Bingyao Huang, Haibin Ling; Proceedings of the IEEE/CVF International Conferenc
e on Computer Vision (ICCV), 2019, pp. 7165-7174
Full projector compensation aims to modify a projector input image such that it
can compensate for both geometric and photometric disturbance of the projection
surface. Traditional methods usually solve the two parts separately, although th
ey are known to correlate with each other. In this paper, we propose the first e
nd-to-end solution, named CompenNet++, to solve the two problems jointly. Our wo
rk non-trivially extends CompenNet, which was recently proposed for photometric
compensation with promising performance. First, we propose a novel geometric cor
rection subnet, which is designed with a cascaded coarse-to-fine structure to le
arn the sampling grid directly from photometric sampling images. Second, by conc
atenating the geometric correction subset with CompenNet, CompenNet++ accomplish

es full projector compensation and is end-to-end trainable. Third, after trainin g, we significantly simplify both geometric and photometric compensation parts, and hence largely improves the running time efficiency. Moreover, we construct t he first setup-independent full compensation benchmark to facilitate the study o n this topic. In our thorough experiments, our method shows clear advantages ove r previous arts with promising compensation quality and meanwhile being practica lly convenient.
****

Deep Parametric Indoor Lighting Estimation
Marc-Andre Gardner, Yannick Hold-Geoffroy, Kalyan Sunkavalli, Christian Gagne , Jean-Francois Lalonde; Proceedings of the IEEE/CVF International Conference o n Computer Vision (ICCV), 2019, pp. 7175-7183
We present a method to estimate lighting from a single image of an indoor scene. Previous work has used an environment map representation that does not account for the localized nature of indoor lighting. Instead, we represent lighting as a set of discrete 3D lights with geometric and photometric parameters. We train a deep neural network to regress these parameters from a single image, on a datas et of environment maps annotated with depth. We propose a differentiable layer t o convert these parameters to an environment map to compute our loss; this bypas ses the challenge of establishing correspondences between estimated and ground t ruth lights. We demonstrate, via quantitative and qualitative evaluations, that our representation and training scheme lead to more accurate results compared to previous work, while allowing for more realistic 3D object compositing with spa tially-varying lighting.
****

FSGAN: Subject Agnostic Face Swapping and Reenactment
Yuval Nirkin, Yosi Keller, Tal Hassner; Proceedings of the IEEE/CVF Internatio nal Conference on Computer Vision (ICCV), 2019, pp. 7184-7193
We present Face Swapping GAN (FSGAN) for face swapping and reenactment. Unlike p revious work, FSGAN is subject agnostic and can be applied to pairs of faces wit hout requiring training on those faces. To this end, we describe a number of tec hnical contributions. We derive a novel recurrent neural network (RNN)-based app roach for face reenactment which adjusts for both pose and expression variations and can be applied to a single image or a video sequence. For video sequences, we introduce continuous interpolation of the face views based on reenactment, De launay Triangulation, and barycentric coordinates. Occluded face regions are han dled by a face completion network. Finally, we use a face blending network for s eamless blending of the two faces while preserving target skin color and lightin g conditions. This network uses a novel Poisson blending loss which combines Poi sson optimization with perceptual loss. We compare our approach to existing stat e-of-the-art systems and show our results to be both qualitatively and quantitat ively superior.
****

Deep Single-Image Portrait Relighting
Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, David W. Jacobs; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7194-720 2
Conventional physically-based methods for relighting portrait images need to sol ve an inverse rendering problem, estimating face geometry, reflectance and light ing. However, the inaccurate estimation of face components can cause strong arti facts in relighting, leading to unsatisfactory results. In this work, we apply a physically-based portrait relighting method to generate a large scale, high qua lity, "in the wild" portrait relighting dataset (DPR). A deep Convolutional Neur al Network (CNN) is then trained using this dataset to generate a relit portrait image by using a source image and a target lighting as input. The training proc edure regularizes the generated results, removing the artifacts caused by physic ally-based relighting methods. A GAN loss is further applied to improve the qual ity of the relit portrait image. Our trained network can relight portrait images with resolutions as high as 1024 x 1024. We evaluate the proposed method on the proposed DPR datset, Flickr portrait dataset and Multi-PIE dataset both qualita

tively and quantitatively. Our experiments demonstrate that the proposed method achieves state-of-the-art results. Please refer to https://zhhoper.github.io/dpr.html for dataset and code.
****

PU-GAN: A Point Cloud Upsampling Adversarial Network
Ruihui Li, Xianzhi Li, Chi-Wing Fu, Daniel Cohen-Or, Pheng-Ann Heng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7203-7212
Point clouds acquired from range scans are often sparse, noisy, and non-uniform. This paper presents a new point cloud upsampling network called PU-GAN, which is formulated based on a generative adversarial network (GAN), to learn a rich variety of point distributions from the latent space and upsample points over patches on object surfaces. To realize a working GAN network, we construct an up-down-up expansion unit in the generator for upsampling point features with error feedback and self-correction, and formulate a self-attention unit to enhance the feature integration. Further, we design a compound loss with adversarial, uniform and reconstruction terms, to encourage the discriminator to learn more latent patterns and enhance the output point distribution uniformity. Qualitative and quantitative evaluations demonstrate the quality of our results over the state-of-the-arts in terms of distribution uniformity, proximity-to-surface, and 3D reconstruction quality.
****

Neural 3D Morphable Models: Spiral Convolutional Networks for 3D Shape Representation Learning and Generation
Giorgos Bouritsas, Sergiy Bokhnyak, Stylianos Ploumpis, Michael Bronstein, Stefanos Zafeiriou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7213-7222
Generative models for 3D geometric data arise in many important applications in 3D computer vision and graphics. In this paper, we focus on 3D deformable shapes that share a common topological structure, such as human faces and bodies. Morphable Models and their variants, despite their linear formulation, have been widely used for shape representation, while most of the recently proposed nonlinear approaches resort to intermediate representations, such as 3D voxel grids or 2D views. In this work, we introduce a novel graph convolutional operator, acting directly on the 3D mesh, that explicitly models the inductive bias of the fixed underlying graph. This is achieved by enforcing consistent local orderings of the vertices of the graph, through the spiral operator, thus breaking the permutation invariance property that is adopted by all the prior work on Graph Neural Networks. Our operator comes by construction with desirable properties (anisotropic, topology-aware, lightweight, easy-to-optimise), and by using it as a building block for traditional deep generative architectures, we demonstrate state-of-the-art results on a variety of 3D shape datasets compared to the linear Morphable Model and other graph convolutional operators.
****

Joint Learning of Saliency Detection and Weakly Supervised Semantic Segmentation
Yu Zeng, Yunzhi Zhuge, Huchuan Lu, Lihe Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7223-7233
Existing weakly supervised semantic segmentation (WSSS) methods usually utilize the results of pre-trained saliency detection (SD) models without explicitly modelling the connections between the two tasks, which is not the most efficient configuration. Here we propose a unified multi-task learning framework to jointly solve WSSS and SD using a single network, i.e. saliency and segmentation network (SSNet). SSNet consists of a segmentation network (SN) and a saliency aggregation module (SAM). For an input image, SN generates the segmentation result and, SAM predicts the saliency of each category and aggregating the segmentation masks of all categories into a saliency map. The proposed network is trained end-to-end with image-level category labels and class-agnostic pixel-level saliency labels. Experiments on PASCAL VOC 2012 segmentation dataset and four saliency benchmark datasets show the performance of our method compares favorably against state-of-the-art weakly supervised segmentation methods and fully supervised saliency

detection methods.
****
Towards High-Resolution Salient Object Detection
Yi Zeng, Pingping Zhang, Jianming Zhang, Zhe Lin, Huchuan Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7234-7243
Deep neural network based methods have made a significant breakthrough in salient object detection. However, they are typically limited to input images with low resolutions (400x400 pixels or less). Little effort has been made to train neural networks to directly handle salient object segmentation in high-resolution images. This paper pushes forward high-resolution saliency detection, and contributes a new dataset, named High-Resolution Salient Object Detection (HRSOD) dataset. To our best knowledge, HRSOD is the first high-resolution saliency detection dataset to date. As another contribution, we also propose a novel approach, which incorporates both global semantic information and local high-resolution details, to address this challenging task. More specifically, our approach consists of a Global Semantic Network (GSN), a Local Refinement Network (LRN) and a Global-Local Fusion Network (GLFN). The GSN extracts the global semantic information based on downsampled entire image. Guided by the results of GSN, the LRN focuses on some local regions and progressively produces high-resolution predictions. The GLFN is further proposed to enforce spatial consistency and boost performance. Experiments illustrate that our method outperforms existing state-of-the-art methods on high-resolution saliency datasets by a large margin, and achieves comparable or even better performance than them on some widely used saliency benchmarks.
****
Event-Based Motion Segmentation by Motion Compensation
Timo Stoffregen, Guillermo Gallego, Tom Drummond, Lindsay Kleeman, Davide Scaramuzza; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7244-7253
In contrast to traditional cameras, whose pixels have a common exposure time, event-based cameras are novel bio-inspired sensors whose pixels work independently and asynchronously output intensity changes (called "events"), with microsecond resolution. Since events are caused by the apparent motion of objects, event-based cameras sample visual information based on the scene dynamics and are, therefore, a more natural fit than traditional cameras to acquire motion, especially at high speeds, where traditional cameras suffer from motion blur. However, distinguishing between events caused by different moving objects and by the camera's ego-motion is a challenging task. We present the first per-event segmentation method for splitting a scene into independently moving objects. Our method jointly estimates the event-object associations (i.e., segmentation) and the motion parameters of the objects (or the background) by maximization of an objective function, which builds upon recent results on event-based motion-compensation. We provide a thorough evaluation of our method on a public dataset, outperforming the state-of-the-art by as much as 10%. We also show the first quantitative evaluation of a segmentation algorithm for event cameras, yielding around 90% accuracy at 4 pixels relative displacement.
****
Depth-Induced Multi-Scale Recurrent Attention Network for Saliency Detection
Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, Huchuan Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7254-7263
In this work, we propose a novel depth-induced multi-scale recurrent attention network for saliency detection. It achieves dramatic performance especially in complex scenarios. There are three main contributions of our network that are experimentally demonstrated to have significant practical merits. First, we design an effective depth refinement block using residual connections to fully extract and fuse multi-level paired complementary cues from RGB and depth streams. Second, depth cues with abundant spatial information are innovatively combined with multi-scale context features for accurately locating salient objects. Third, we bo

ost our model's performance by a novel recurrent attention module inspired by In
ternal Generative Mechanism of human brain. This module can generate more accura
te saliency results via comprehensively learning the internal semantic relation
of the fused feature and progressively optimizing local details with memory-orie
nted scene understanding. In addition, we create a large scale RGB-D dataset con
taining more complex scenarios, which can contribute to comprehensively evaluati
ng saliency models. Extensive experiments on six public datasets and ours demons
trate that our method can accurately identify salient objects and achieve consis
tently superior performance over 16 state-of-the-art RGB and RGB-D approaches.
****

Stacked Cross Refinement Network for Edge-Aware Salient Object Detection
Zhe Wu, Li Su, Qingming Huang; Proceedings of the IEEE/CVF International Confe
rence on Computer Vision (ICCV), 2019, pp. 7264-7273
Salient object detection is a fundamental computer vision task. The majority of
existing algorithms focus on aggregating multi-level features of pre-trained con
volutional neural networks. Moreover, some researchers attempt to utilize edge i
nformation for auxiliary training. However, existing edge-aware models design un
idirectional frameworks which only use edge features to improve the segmentation
features. Motivated by the logical interrelations between binary segmentation a
nd edge maps, we propose a novel Stacked Cross Refinement Network (SCRN) for sal
ient object detection in this paper. Our framework aims to simultaneously refine
multi-level features of salient object detection and edge detection by stacking
Cross Refinement Unit (CRU). According to the logical interrelations, the CRU d
esigns two direction-specific integration operations, and bidirectionally passes
messages between the two tasks. Incorporating the refined edge-preserving featu
res with the typical U-Net, our model detects salient objects accurately. Extens
ive experiments conducted on six benchmark datasets demonstrate that our method
outperforms existing state-of-the-art algorithms in both accuracy and efficiency
. Besides, the attribute-based performance on the SOC dataset show that the prop
osed model ranks first in the majority of challenging scenes. Code can be found
at https://github.com/wuzhe71/SCAN.
****

Motion Guided Attention for Video Salient Object Detection
Haofeng Li, Guanqi Chen, Guanbin Li, Yizhou Yu; Proceedings of the IEEE/CVF I
nternational Conference on Computer Vision (ICCV), 2019, pp. 7274-7283
Video salient object detection aims at discovering the most visually distinctive
objects in a video. How to effectively take object motion into consideration du
ring video salient object detection is a critical issue. Existing state-of-the-a
rt methods either do not explicitly model and harvest motion cues or ignore spat
ial contexts within optical flow images. In this paper, we develop a multi-task
motion guided video salient object detection network, which learns to accomplish
two sub-tasks using two sub-networks, one sub-network for salient object detect
ion in still images and the other for motion saliency detection in optical flow
images. We further introduce a series of novel motion guided attention modules,
which utilize the motion saliency sub-network to attend and enhance the sub-netw
ork for still images. These two sub-networks learn to adapt to each other by end
-to-end training. Experimental results demonstrate that the proposed method sign
ificantly outperforms existing state-of-the-art algorithms on a wide range of be
nchmarks. We hope our simple and effective approach will serve as a solid baseli
ne and help ease future research in video salient object detection. Code and mod
els will be made available.
****

Semi-Supervised Video Salient Object Detection Using Pseudo-Labels
Pengxiang Yan, Guanbin Li, Yuan Xie, Zhen Li, Chuan Wang, Tianshui Chen, L
iang Lin; Proceedings of the IEEE/CVF International Conference on Computer Visio
n (ICCV), 2019, pp. 7284-7293
Deep learning-based video salient object detection has recently achieved great s
uccess with its performance significantly outperforming any other unsupervised m
ethods. However, existing data-driven approaches heavily rely on a large quantit
y of pixel-wise annotated video frames to deliver such promising results. In thi

s paper, we address the semi-supervised video salient object detection task usin g pseudo-labels. Specifically, we present an effective video saliency detector t hat consists of a spatial refinement network and a spatiotemporal module. Based on the same refinement network and motion information in terms of optical flow, we further propose a novel method for generating pixel-level pseudo-labels from sparsely annotated frames. By utilizing the generated pseudo-labels together wit h a part of manual annotations, our video saliency detector learns spatial and t emporal cues for both contrast inference and coherence enhancement, thus produci ng accurate saliency maps. Experimental results demonstrate that our proposed se mi-supervised method even greatly outperforms all the state-of-the-art fully sup ervised methods across three public benchmarks of VOS, DAVIS, and FBMS.
****

Joint Learning of Semantic Alignment and Object Landmark Detection
Sangryul Jeon, Dongbo Min, Seungryong Kim, Kwanghoon Sohn; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7294-730 3
Convolutional neural networks (CNNs) based approaches for semantic alignment and object landmark detection have improved their performance significantly. Curren t efforts for the two tasks focus on addressing the lack of massive training dat a through weakly- or unsupervised learning frameworks. In this paper, we present a joint learning approach for obtaining dense correspondences and discovering o bject landmarks from semantically similar images. Based on the key insight that the two tasks can mutually provide supervisions to each other, our networks acco mplish this through a joint loss function that alternatively imposes a consisten cy constraint between the two tasks, thereby boosting the performance and addres sing the lack of training data in a principled manner. To the best of our knowle dge, this is the first attempt to address the lack of training data for the two tasks through the joint learning. To further improve the robustness of our frame work, we introduce a probabilistic learning formulation that allows only reliabl e matches to be used in the joint learning process. With the proposed method, st ate-of-the-art performance is attained on several benchmarks for semantic matchi ng and landmark detection.
****

RainFlow: Optical Flow Under Rain Streaks and Rain Veiling Effect
Ruoteng Li, Robby T. Tan, Loong-Fah Cheong, Angelica I. Aviles-Rivero, Qingn an Fan, Carola-Bibiane Schonlieb; Proceedings of the IEEE/CVF International Con ference on Computer Vision (ICCV), 2019, pp. 7304-7313
Optical flow in heavy rainy scenes is challenging due to the presence of both ra in steaks and rain veiling effect, which break the existing optical flow constra ints. Concerning this, we propose a deep-learning based optical flow method desi gned to handle heavy rain. We introduce a feature multiplier in our network that transforms the features of an image affected by the rain veiling effect into fe atures that are less affected by it, which we call veiling-invariant features. W e establish a new mapping operation in the feature space to produce streak-invar iant features. The operation is based on a feature pyramid structure of the inpu t images, and the basic idea is to preserve the chromatic features of the backgr ound scenes while canceling the rain-streak patterns. Both the veiling-invariant and streak-invariant features are computed and optimized automatically based on the the accuracy of our optical flow estimation. Our network is end-to-end, and handles both rain streaks and the veiling effect in an integrated framework. Ex tensive experiments show the effectiveness of our method, which outperforms the state of the art method and other baseline methods. We also show that our networ k can robustly maintain good performance on clean (no rain) images even though i t is trained under rain image data.
****

GridDehazeNet: Attention-Based Multi-Scale Network for Image Dehazing
Xiaohong Liu, Yongrui Ma, Zhihao Shi, Jun Chen; Proceedings of the IEEE/CVF I nternational Conference on Computer Vision (ICCV), 2019, pp. 7314-7323
We propose an end-to-end trainable Convolutional Neural Network (CNN), named Gri dDehazeNet, for single image dehazing. The GridDehazeNet consists of three modul

es: pre-processing, backbone, and post-processing. The trainable pre-processing module can generate learned inputs with better diversity and more pertinent features as compared to those derived inputs produced by hand-selected pre-processing methods. The backbone module implements a novel attention-based multi-scale estimation on a grid network, which can effectively alleviate the bottleneck issue often encountered in the conventional multi-scale approach. The post-processing module helps to reduce the artifacts in the final output. Experimental results indicate that the GridDehazeNet outperforms the state-of-the-arts on both synthetic and real-world images. The proposed hazing method does not rely on the atmosphere scattering model, and we provide an explanation as to why it is not necessarily beneficial to take advantage of the dimension reduction offered by the atmosphere scattering model for image dehazing, even if only the dehazing results on synthetic images are concerned.
****

Learning to See Moving Objects in the Dark
Haiyang Jiang, Yinqiang Zheng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7324-7333
Video surveillance systems have wide range of utilities, yet easily suffer from great quality degeneration under dim light circumstances. Industrial solutions mainly use extra near-infrared illuminations, even though it doesn't preserve color and texture information. A variety of researches enhanced low-light videos shot by visible light cameras, while they either relied on task specific preconditions or trained with synthetic datasets. We propose a novel optical system to capture bright and dark videos of the exact same scenes, generating training and groud truth pairs for authentic low-light video dataset. A fully convolutional network with 3D and 2D miscellaneous operations is utilized to learn an enhancement mapping with proper spatial-temporal transformation from raw camera sensor data to bright RGB videos. Experiments show promising results by our method, and it outperforms state-of-the-art low-light image/video enhancement algorithms.
****

SegSort: Segmentation by Discriminative Sorting of Segments
Jyh-Jing Hwang, Stella X. Yu, Jianbo Shi, Maxwell D. Collins, Tien-Ju Yang, Xiao Zhang, Liang-Chieh Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7334-7344
Almost all existing deep learning approaches for semantic segmentation tackle this task as a pixel-wise classification problem. Yet humans understand a scene not in terms of pixels, but by decomposing it into perceptual groups and structures that are the basic building blocks of recognition. This motivates us to propose an end-to-end pixel-wise metric learning approach that mimics this process. In our approach, the optimal visual representation determines the right segmentation within individual images and associates segments with the same semantic classes across images. The core visual learning problem is therefore to maximize the similarity within segments and minimize the similarity between segments. Given a model trained this way, inference is performed consistently by extracting pixel-wise embeddings and clustering, with the semantic label determined by the majority vote of its nearest neighbors from an annotated set. As a result, we present the SegSort, as a first attempt using deep learning for unsupervised semantic segmentation, achieving 76% performance of its supervised counterpart. When supervision is available, SegSort shows consistent improvements over conventional approaches based on pixel-wise softmax training. Additionally, our approach produces more precise boundaries and consistent region predictions. The proposed SegSort further produces an interpretable result, as each choice of label can be easily understood from the retrieved nearest segments.
****

What Synthesis Is Missing: Depth Adaptation Integrated With Weak Supervision for Indoor Scene Parsing
Keng-Chi Liu, Yi-Ting Shen, Jan P. Klopp, Liang-Gee Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7345-7354
Scene Parsing is a crucial step to enable autonomous systems to understand and interact with their surroundings. Supervised deep learning methods have made grea

t progress in solving scene parsing problems, however, come at the cost of labor ious manual pixel-level annotation. Synthetic data as well as weak supervision h ave been investigated to alleviate this effort. Nonetheless, synthetically gener ated data still suffers from severe domain shift while weak labels often lack pr ecision. Moreover, most existing works for weakly supervised scene parsing are l imited to salient foreground objects. The aim of this work is hence twofold: Exp loit synthetic data where feasible and integrate weak supervision where necessar y. More concretely, we address this goal by utilizing depth as transfer domain b ecause its synthetic-to-real discrepancy is much lower than for color. At the sa me time, we perform weak localization from easily obtainable image level labels and integrate both using a novel contour-based scheme. Our approach is implement ed as a teacher-student learning framework to solve the transfer learning proble m by generating a pseudo ground truth. Using only depth-based adaptation, this a pproach already outperforms previous transfer learning approaches on the popular indoor scene parsing SUN RGB-D dataset. Our proposed two-stage integration more than halves the gap towards fully supervised methods when compared to previous state-of-the-art in transfer learning.

****

AdaptIS: Adaptive Instance Selection Network

Konstantin Sofiiuk, Olga Barinova, Anton Konushin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7355-7363

We present Adaptive Instance Selection network architecture for class-agnostic i nstance segmentation. Given an input image and a point (x, y), it generates a ma sk for the object located at (x, y). The network adapts to the input point with a help of AdaIN layers [??], thus producing different masks for different object s on the same image. AdaptIS generates pixel-accurate object masks, therefore it accurately segments objects of complex shape or severely occluded ones. AdaptIS can be easily combined with standard semantic segmentation pipeline to perform panoptic segmentation. To illustrate the idea, we perform experiments on a chall enging toy problem with difficult occlusions. Then we extensively evaluate the m ethod on panoptic segmentation benchmarks. We obtain state-of-the-art results on Cityscapes and Mapillary even without pretraining on COCO, and show competitive results on a challenging COCO dataset. The source code of the method and the tr ained models are available at https://github.com/saic-vul/adaptis https://gith ub.com/saic-vul/adaptis .

****

DADA: Depth-Aware Domain Adaptation in Semantic Segmentation

Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, Patrick Perez; Pr oceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 20 19, pp. 7364-7373

Unsupervised domain adaptation (UDA) is important for applications where large s cale annotation of representative data is challenging. For semantic segmentation in particular, it helps deploy on real "target domain" data models that are tra ined on annotated images from a different "source domain", notably a virtual env ironment. To this end, most previous works consider semantic segmentation as the only mode of supervision for source domain data, while ignoring other, possibly available, information like depth. In this work, we aim at exploiting at best s uch a privileged information while training the UDA model. We propose a unified depth-aware UDA framework that leverages in several complementary ways the knowl edge of dense depth in the source domain. As a result, the performance of the tr ained semantic segmentation model on the target domain is boosted. Our novel app roach indeed achieves state-of-the-art performance on different challenging synt hetic-2-real benchmarks.

****

Guided Curriculum Model Adaptation and Uncertainty-Aware Evaluation for Semantic Nighttime Image Segmentation

Christos Sakaridis, Dengxin Dai, Luc Van Gool; Proceedings of the IEEE/CVF Int ernational Conference on Computer Vision (ICCV), 2019, pp. 7374-7383

Most progress in semantic segmentation reports on daytime images taken under fav orable illumination conditions. We instead address the problem of semantic segme

ntation of nighttime images and improve the state-of-the-art, by adapting daytime models to nighttime without using nighttime annotations. Moreover, we design a new evaluation framework to address the substantial uncertainty of semantics in nighttime images. Our central contributions are: 1) a curriculum framework to gradually adapt semantic segmentation models from day to night via labeled synthetic images and unlabeled real images, both for progressively darker times of day, which exploits cross-time-of-day correspondences for the real images to guide the inference of their labels; 2) a novel uncertainty-aware annotation and evaluation framework and metric for semantic segmentation, designed for adverse conditions and including image regions beyond human recognition capability in the evaluation in a principled fashion; 3) the Dark Zurich dataset, which comprises 2416 unlabeled nighttime and 2920 unlabeled twilight images with correspondences to their daytime counterparts plus a set of 151 nighttime images with fine pixel-level annotations created with our protocol, which serves as a first benchmark to perform our novel evaluation. Experiments show that our guided curriculum adaptation significantly outperforms state-of-the-art methods on real nighttime sets both for standard metrics and our uncertainty-aware metric. Furthermore, our uncertainty-aware evaluation reveals that selective invalidation of predictions can lead to better results on data with ambiguous content such as our nighttime benchmark and profit safety-oriented applications which involve invalid inputs.
****

SceneGraphNet: Neural Message Passing for 3D Indoor Scene Augmentation
Yang Zhou, Zachary While, Evangelos Kalogerakis; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7384-7392
In this paper we propose a neural message passing approach to augment an input 3D indoor scene with new objects matching their surroundings. Given an input, potentially incomplete, 3D scene and a query location, our method predicts a probability distribution over object types that fit well in that location. Our distribution is predicted though passing learned messages in a dense graph whose nodes represent objects in the input scene and edges represent spatial and structural relationships. By weighting messages through an attention mechanism, our method learns to focus on the most relevant surrounding scene context to predict new scene objects. We found that our method significantly outperforms state-of-the-art approaches in terms of correctly predicting objects missing in a scene based on our experiments in the SUNCG dataset. We also demonstrate other applications of our method, including context-based 3D object recognition and iterative scene generation.
****

SkyScapes Fine-Grained Semantic Understanding of Aerial Scenes
Seyed Majid Azimi, Corentin Henry, Lars Sommer, Arne Schumann, Eleonora Vig; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7393-7403
Understanding the complex urban infrastructure with centimeter-level accuracy is essential for many applications from autonomous driving to mapping, infrastructure monitoring, and urban management. Aerial images provide valuable information over a large area instantaneously; nevertheless, no current dataset captures the complexity of aerial scenes at the level of granularity required by real-world applications. To address this, we introduce SkyScapes, an aerial image dataset with highly-accurate, fine-grained annotations for pixel-level semantic labeling. SkyScapes provides annotations for 31 semantic categories ranging from large structures, such as buildings, roads and vegetation, to fine details, such as 12 (sub-)categories of lane markings. We have defined two main tasks on this dataset: dense semantic segmentation and multi-class lane-marking prediction. We carry out extensive experiments to evaluate state-of-the-art segmentation methods on SkyScapes. Existing methods struggle to deal with the wide range of classes, object sizes, scales, and fine details present. We therefore propose a novel multi-task model, which incorporates semantic edge detection and is better tuned for feature extraction from a wide range of scales. This model achieves notable improvements over the baselines in region outlines and level of detail on both tasks.
****

Transferable Representation Learning in Vision-and-Language Navigation
Haoshuo Huang, Vihan Jain, Harsh Mehta, Alexander Ku, Gabriel Magalhaes, Jason Baldridge, Eugene Ie; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7404-7413
Vision-and-Language Navigation (VLN) tasks such as Room-to-Room (R2R) require machine agents to interpret natural language instructions and learn to act in visually realistic environments to achieve navigation goals. The overall task requires competence in several perception problems: successful agents combine spatio-temporal, vision and language understanding to produce appropriate action sequences. Our approach adapts pre-trained vision and language representations to relevant in-domain tasks making them more effective for VLN. Specifically, the representations are adapted to solve both a cross-modal sequence alignment and sequence coherence task. In the sequence alignment task, the model determines whether an instruction corresponds to a sequence of visual frames. In the sequence coherence task, the model determines whether the perceptual sequences are predictive sequentially in the instruction-conditioned latent space. By transferring the domain-adapted representations, we improve competitive agents in R2R as measured by the success rate weighted by path length (SPL) metric.
****
Towards Unsupervised Image Captioning With Shared Multimodal Embeddings
Iro Laina, Christian Rupprecht, Nassir Navab; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7414-7424
Understanding images without explicit supervision has become an important problem in computer vision. In this paper, we address image captioning by generating language descriptions of scenes without learning from annotated pairs of images and their captions. The core component of our approach is a shared latent space that is structured by visual concepts. In this space, the two modalities should be indistinguishable. A language model is first trained to encode sentences into semantically structured embeddings. Image features that are translated into this embedding space can be decoded into descriptions through the same language model, similarly to sentence embeddings. This translation is learned from weakly paired images and text using a loss robust to noisy assignments and a conditional adversarial component. Our approach allows to exploit large text corpora outside the annotated distributions of image/caption data. Our experiments show that the proposed domain alignment learns a semantically meaningful representation which outperforms previous work.
****
ViCo: Word Embeddings From Visual Co-Occurrences
Tanmay Gupta, Alexander Schwing, Derek Hoiem; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7425-7434
We propose to learn word embeddings from visual co-occurrences. Two words co-occur visually if both words apply to the same image or image region. Specifically, we extract four types of visual co-occurrences between object and attribute words from large-scale, textually-annotated visual databases like VisualGenome and ImageNet. We then train a multi-task log-bilinear model that compactly encodes word "meanings" represented by each co-occurrence type into a single visual word-vector. Through unsupervised clustering, supervised partitioning, and a zero-shot-like generalization analysis we show that our word embeddings complement text-only embeddings like GloVe by better representing similarities and differences between visual concepts that are difficult to obtain from text corpora alone. We further evaluate our embeddings on five downstream applications, four of which are vision-language tasks. Augmenting GloVe with our embeddings yields gains on all tasks. We also find that random embeddings perform comparably to learned embeddings on all supervised vision-language tasks, contrary to conventional wisdom.
****
Seq-SG2SL: Inferring Semantic Layout From Scene Graph Through Sequence to Sequence Learning
Boren Li, Boyu Zhuang, Mingyang Li, Jian Gu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7435-7443
Generating semantic layout from scene graph is a crucial intermediate task conne

cting text to image. We present a conceptually simple, flexible and general framework using sequence to sequence (seq-to-seq) learning for this task. The framework, called Seq-SG2SL, derives sequence proxies for the two modality and a Transformer-based seq-to-seq model learns to transduce one into the other. A scene graph is decomposed into a sequence of semantic fragments (SF), one for each relationship. A semantic layout is represented as the consequence from a series of brick-action code segments (BACS), dictating the position and scale of each object bounding box in the layout. Viewing the two building blocks, SF and BACS, as corresponding terms in two different vocabularies, a seq-to-seq model is fittingly used to translate. A new metric, semantic layout evaluation understudy (SLEU), is devised to evaluate the task of semantic layout prediction inspired by BLEU. SLEU defines relationships within a layout as unigrams and looks at the spatial distribution for n-grams. Unlike the binary precision of BLEU, SLEU allows for some tolerances spatially through thresholding the Jaccard Index and is consequently more adapted to the task. Experimental results on the challenging Visual Genome dataset show improvement over a non-sequential approach based on graph convolution.
****

U-CAM: Visual Explanation Using Uncertainty Based Class Activation Maps
Badri N. Patro, Mayank Lunayach, Shivansh Patel, Vinay P. Namboodiri; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7444-7453
Understanding and explaining deep learning models is an imperative task. Towards this, we propose a method that obtains gradient-based certainty estimates that also provide visual attention maps. Particularly, we solve for visual question answering task. We incorporate modern probabilistic deep learning methods that we further improve by using the gradients for these estimates. These have two-fold benefits: a) improvement in obtaining the certainty estimates that correlate better with misclassified samples and b) improved attention maps that provide state-of-the-art results in terms of correlation with human attention regions. The improved attention maps result in consistent improvement for various methods for visual question answering. Therefore, the proposed technique can be thought of as a recipe for obtaining improved certainty estimates and explanation for deep learning models. We provide detailed empirical analysis for the visual question answering task on all standard benchmarks and comparison with state of the art methods.
****

See-Through-Text Grouping for Referring Image Segmentation
Ding-Jie Chen, Songhao Jia, Yi-Chen Lo, Hwann-Tzong Chen, Tyng-Luh Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7454-7463
Motivated by the conventional grouping techniques to image segmentation, we develop their DNN counterpart to tackle the referring variant. The proposed method is driven by a convolutional-recurrent neural network (ConvRNN) that iteratively carries out top-down processing of bottom-up segmentation cues. Given a natural language referring expression, our method learns to predict its relevance to each pixel and derives a See-through-Text Embedding Pixelwise (STEP) heatmap, which reveals segmentation cues of pixel level via the learned visual-textual co-embedding. The ConvRNN performs a top-down approximation by converting the STEP heatmap into a refined one, whereas the improvement is expected from training the network with a classification loss from the ground truth. With the refined heatmap, we update the textual representation of the referring expression by re-evaluating its attention distribution and then compute a new STEP heatmap as the next input to the ConvRNN. Boosting by such collaborative learning, the framework can progressively and simultaneously yield the desired referring segmentation and reasonable attention distribution over the referring sentence. Our method is general and does not rely on, say, the outcomes of object detection from other DNN models, while achieving state-of-the-art performance in all of the four datasets in the experiments.
****

VideoBERT: A Joint Model for Video and Language Representation Learning
Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, Cordelia Schmid; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7464-7473
Self-supervised learning has become increasingly important to leverage the abundance of unlabeled data available on platforms like YouTube. Whereas most existing approaches learn low-level representations, we propose a joint visual-linguistic model to learn high-level features without any explicit supervision. In particular, inspired by its recent success in language modeling, we build upon the BERT model to learn bidirectional joint distributions over sequences of visual and linguistic tokens, derived from vector quantization of video data and off-the-shelf speech recognition outputs, respectively. We use VideoBERT in numerous tasks, including action classification and video captioning. We show that it can be applied directly to open-vocabulary classification, and confirm that large amounts of training data and cross-modal information are critical to performance. Furthermore, we outperform the state-of-the-art on video captioning, and quantitative results verify that the model learns high-level semantic features.
****

Language Features Matter: Effective Language Representations for Vision-Language Tasks
Andrea Burns, Reuben Tan, Kate Saenko, Stan Sclaroff, Bryan A. Plummer; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7474-7483
Shouldn't language and vision features be treated equally in vision-language (VL) tasks? Many VL approaches treat the language component as an afterthought, using simple language models that are either built upon fixed word embeddings trained on text-only data or are learned from scratch. We conclude that language features deserve more attention, which has been informed by experiments which compare different word embeddings, language models, and embedding augmentation steps on five common VL tasks: image-sentence retrieval, image captioning, visual question answering, phrase grounding, and text-to-clip retrieval. Our experiments provide some striking results; an average embedding language model outperforms a LSTM on retrieval-style tasks; state-of-the-art representations such as BERT perform relatively poorly on vision-language tasks. From this comprehensive set of experiments we can propose a set of best practices for incorporating the language component of vision-language tasks. To further elevate language features, we also show that knowledge in vision-language problems can be transferred across tasks to gain performance with multi-task training. This multi-task training is applied to a new Graph Oriented Vision-Language Embedding (GrOVLE), which we adapt from Word2Vec using WordNet and an original visual-language graph built from Visual Genome, providing a ready-to-use vision-language embedding: http://ai.bu.edu/grovle.
****

Semantic Stereo Matching With Pyramid Cost Volumes
Zhenyao Wu, Xinyi Wu, Xiaoping Zhang, Song Wang, Lili Ju; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7484-7493
The accuracy of stereo matching has been greatly improved by using deep learning with convolutional neural networks. To further capture the details of disparity maps, in this paper, we propose a novel semantic stereo network named SSPCV-Net, which includes newly designed pyramid cost volumes for describing semantic and spatial information on multiple levels. The semantic features are inferred by a semantic segmentation subnetwork while the spatial features are derived by hierarchical spatial pooling. In the end, we design a 3D multi-cost aggregation module to integrate the extracted multilevel features and perform regression for accurate disparity maps. We conduct comprehensive experiments and comparisons with some recent stereo matching networks on Scene Flow, KITTI 2015 and 2012, and Cityscapes benchmark datasets, and the results show that the proposed SSPCV-Net significantly promotes the state-of-the-art stereo-matching performance.
****

Learning Relationships for Multi-View 3D Object Recognition
Ze Yang, Liwei Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7505-7514
Recognizing 3D object has attracted plenty of attention recently, and view-based methods have achieved best results until now. However, previous view-based methods ignore the region-to-region and view-to-view relationships between different view images, which are crucial for multi-view 3D object representation. To tackle this problem, we propose a Relation Network to effectively connect corresponding regions from different viewpoints, and therefore reinforce the information of individual view image. In addition, the Relation Network exploits the inter-relationships over a group of views, and integrates those views to obtain a discriminative 3D object representation. Systematic experiments conducted on ModelNet dataset demonstrate the effectiveness of our proposed methods for both 3D object recognition and retrieval tasks.
****

View N-Gram Network for 3D Object Retrieval
Xinwei He, Tengteng Huang, Song Bai, Xiang Bai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7515-7524
How to aggregate multi-view representations of a 3D object into an informative and discriminative one remains a key challenge for multi-view 3D object retrieval. Existing methods either use view-wise pooling strategies which neglect the spatial information across different views or employ recurrent neural networks which may face the efficiency problem. To address these issues, we propose an effective and efficient framework called View N-gram Network (VNN). Inspired by n-gram models in natural language processing, VNN divides the view sequence into a set of visual n-grams, which involve overlapping consecutive view sub-sequences. By doing so, spatial information across multiple views is captured, which helps to learn a discriminative global embedding for each 3D object. Experiments on 3D shape retrieval benchmarks, including ModelNet10, ModelNet40 and ShapeNetCore55 datasets, demonstrate the superiority of our proposed method.
****

Expert Sample Consensus Applied to Camera Re-Localization
Eric Brachmann, Carsten Rother; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7525-7534
Fitting model parameters to a set of noisy data points is a common problem in computer vision. In this work, we fit the 6D camera pose to a set of noisy correspondences between the 2D input image and a known 3D environment. We estimate these correspondences from the image using a neural network. Since the correspondences often contain outliers, we utilize a robust estimator such as Random Sample Consensus (RANSAC) or Differentiable RANSAC (DSAC) to fit the pose parameters. When the problem domain, e.g. the space of all 2D-3D correspondences, is large or ambiguous, a single network does not cover the domain well. Mixture of Experts (MoE) is a popular strategy to divide a problem domain among an ensemble of specialized networks, so called experts, where a gating network decides which expert is responsible for a given input. In this work, we introduce Expert Sample Consensus (ESAC), which integrates DSAC in a MoE. Our main technical contribution is an efficient method to train ESAC jointly and end-to-end. We demonstrate experimentally that ESAC handles two real-world problems better than competing methods, i.e. scalability and ambiguity. We apply ESAC to fitting simple geometric models to synthetic images, and to camera re-localization for difficult, real datasets.
****

Semantic Part Detection via Matching: Learning to Generalize to Novel Viewpoints From Limited Training Data
Yutong Bai, Qing Liu, Lingxi Xie, Weichao Qiu, Yan Zheng, Alan L. Yuille; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7535-7545
Detecting semantic parts of an object is a challenging task, particularly because it is hard to annotate semantic parts and construct large datasets. In this paper, we present an approach which can learn from a small annotated dataset conta

ining a limited range of viewpoints and generalize to detect semantic parts for a much larger range of viewpoints. The approach is based on our matching algorithm, which is used for finding accurate spatial correspondence between two images and transplanting semantic parts annotated on one image to the other. Images in the training set are matched to synthetic images rendered from a 3D CAD model, following which a clustering algorithm is used to automatically annotate semantic parts of the CAD model. During the testing period, this CAD model can synthesize annotated images under every viewpoint. These synthesized images are matched to images in the testing set to detect semantic parts in novel viewpoints. Our algorithm is simple, intuitive, and contains very few parameters. Experiments show our method outperforms standard deep learning approaches and, in particular, performs much better on novel viewpoints. For facilitating the future research, code is available: https://github.com/ytongbai/SemanticPartDetection

****

Dynamic Points Agglomeration for Hierarchical Point Sets Learning

Jinxian Liu, Bingbing Ni, Caiyuan Li, Jiancheng Yang, Qi Tian; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7546-7555

Many previous works on point sets learning achieve excellent performance with hierarchical architecture. Their strategies towards points agglomeration, however, only perform points sampling and grouping in original Euclidean space in a fixed way. These heuristic and task-irrelevant strategies severely limit their ability to adapt to more varied scenarios. To this end, we develop a novel hierarchical point sets learning architecture, with dynamic points agglomeration. By exploiting the relation of points in semantic space, a module based on graph convolution network is designed to learn a soft points cluster agglomeration. We construct a hierarchical architecture that gradually agglomerates points by stacking this learnable and lightweight module. In contrast to fixed points agglomeration strategy, our method can handle more diverse situations robustly and efficiently. Moreover, we propose a parameter sharing scheme for reducing memory usage and computational burden induced by the agglomeration module. Extensive experimental results on several point cloud analytic tasks, including classification and segmentation, well demonstrate the superior performance of our dynamic hierarchical learning framework over current state-of-the-art methods.

****

Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints

Ning Yu, Larry S. Davis, Mario Fritz; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7556-7566

Recent advances in Generative Adversarial Networks (GANs) have shown increasing success in generating photorealistic images. But they also raise challenges to visual forensics and model attribution. We present the first study of learning GAN fingerprints towards image attribution and using them to classify an image as real or GAN-generated. For GAN-generated images, we further identify their sources. Our experiments show that (1) GANs carry distinct model fingerprints and leave stable fingerprints in their generated images, which support image attribution; (2) even minor differences in GAN training can result in different fingerprints, which enables fine-grained model authentication; (3) fingerprints persist across different image frequencies and patches and are not biased by GAN artifacts; (4) fingerprint finetuning is effective in immunizing against five types of adversarial image perturbations; and (5) comparisons also show our learned fingerprints consistently outperform several baselines in a variety of setups.

****

Dual Adversarial Inference for Text-to-Image Synthesis

Qicheng Lao, Mohammad Havaei, Ahmad Pesaranghader, Francis Dutil, Lisa Di Jorio, Thomas Fevens; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7567-7576

Synthesizing images from a given text description involves engaging two types of information: the content, which includes information explicitly described in the text (e.g., color, composition, etc.), and the style, which is usually not well described in the text (e.g., location, quantity, size, etc.). However, in prev

ious works, it is typically treated as a process of generating images only from the content, i.e., without considering learning meaningful style representations . In this paper, we aim to learn two variables that are disentangled in the late nt space, representing content and style respectively. We achieve this by augmen ting current text-to-image synthesis frameworks with a dual adversarial inferenc e mechanism. Through extensive experiments, we show that our model learns, in an unsupervised manner, style representations corresponding to certain meaningful information present in the image that are not well described in the text. The ne w framework also improves the quality of synthesized images when evaluated on Ox ford-102, CUB and COCO datasets.

****

View-LSTM: Novel-View Video Synthesis Through View Decomposition

Mohamed Ilyes Lakhal, Oswald Lanz, Andrea Cavallaro; Proceedings of the IEEE/C VF International Conference on Computer Vision (ICCV), 2019, pp. 7577-7587

We tackle the problem of synthesizing a video of multiple moving people as seen from a novel view, given only an input video and depth information or human pose s of the novel view as prior. This problem requires a model that learns to trans form input features into target features while maintaining temporal consistency. To this end, we learn an invariant feature from the input video that is shared across all viewpoints of the same scene and a view-dependent feature obtained us ing the target priors. The proposed approach, View-LSTM, is a recurrent neural n etwork structure that accounts for the temporal consistency and target feature a pproximation constraints. We validate View-LSTM by designing an end-to-end gener ator for novel-view video synthesis. Experiments on a large multi-view action re cognition dataset validate the proposed model.

****

HoloGAN: Unsupervised Learning of 3D Representations From Natural Images

Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, Yong-Liang Yang ; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) , 2019, pp. 7588-7597

We propose a novel generative adversarial network (GAN) for the task of unsuperv ised learning of 3D representations from natural images. Most generative models rely on 2D kernels to generate images and make few assumptions about the 3D worl d. These models therefore tend to create blurry images or artefacts in tasks tha t require a strong 3D understanding, such as novel-view synthesis. HoloGAN inste ad learns a 3D representation of the world, and to render this representation in a realistic manner. Unlike other GANs, HoloGAN provides explicit control over t he pose of generated objects through rigid-body transformations of the learnt 3D features. Our experiments show that using explicit 3D features enables HoloGAN to disentangle 3D pose and identity, which is further decomposed into shape and appearance, while still being able to generate images with similar or higher vis ual quality than other generative models. HoloGAN can be trained end-to-end from unlabelled 2D images only. Particularly, we do not require pose labels, 3D shap es, or multiple views of the same objects. This shows that HoloGAN is the first generative model that learns 3D representations from natural images in an entire ly unsupervised manner.

****

Unpaired Image-to-Speech Synthesis With Multimodal Information Bottleneck

Shuang Ma, Daniel McDuff, Yale Song; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7598-7607

Deep generative models have led to significant advances in cross-modal generatio n such as text-to-image synthesis. Training these models typically requires pair ed data with direct correspondence between modalities. We introduce the novel pr oblem of translating instances from one modality to another without paired data by leveraging an intermediate modality shared by the two other modalities. To de monstrate this, we take the problem of translating images to speech. In this cas e, one could leverage disjoint datasets with one shared modality, e.g., image-te xt pairs and text-speech pairs, with text as the shared modality. We call this p roblem "skip-modal generation" because the shared modality is skipped during the generation process. We propose a multimodal information bottleneck approach tha

t learns the correspondence between modalities from unpaired data (image and speech) by leveraging the shared modality (text). We address fundamental challenges of skip-modal generation: 1) learning multimodal representations using a single model, 2) bridging the domain gap between two unrelated datasets, and 3) learning the correspondence between modalities from unpaired data. We show qualitative results on image-to-speech synthesis; this is the first time such results have been reported in the literature. We also show that our approach improves performance on traditional cross-modal generation, suggesting that it improves data efficiency in solving individual tasks.
****

Improved Conditional VRNNs for Video Prediction
Lluis Castrejon, Nicolas Ballas, Aaron Courville; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7608-7617
Predicting future frames for a video sequence is a challenging generative modeling task. Promising approaches include probabilistic latent variable models such as the Variational Auto-Encoder. While VAEs can handle uncertainty and model multiple possible future outcomes, they have a tendency to produce blurry predictions. In this work we argue that this is a sign of underfitting. To address this issue, we propose to increase the expressiveness of the latent distributions and to use higher capacity likelihood models. Our approach relies on a hierarchy of latent variables, which defines a family of flexible prior and posterior distributions in order to better model the probability of future sequences. We validate our proposal through a series of ablation experiments and compare our approach to current state-of-the-art latent variable models. Our method performs favorably under several metrics in three different datasets.
****

Visualizing the Invisible: Occluded Vehicle Segmentation and Recovery
Xiaosheng Yan, Feigege Wang, Wenxi Liu, Yuanlong Yu, Shengfeng He, Jia Pan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7618-7627
In this paper, we propose a novel iterative multi-task framework to complete the segmentation mask of an occluded vehicle and recover the appearance of its invisible parts. In particular, firstly, to improve the quality of the segmentation completion, we present two coupled discriminators that introduce an auxiliary 3D model pool for sampling authentic silhouettes as adversarial samples. In addition, we propose a two-path structure with a shared network to enhance the appearance recovery capability. By iteratively performing the segmentation completion and the appearance recovery, the results will be progressively refined. To evaluate our method, we present a dataset, Occluded Vehicle dataset, containing synthetic and real-world occluded vehicle images. Based on this dataset, we conduct comparison experiments and demonstrate that our model outperforms the state-of-the-arts in both tasks of recovering segmentation mask and appearance for occluded vehicles. Moreover, we also demonstrate that our appearance recovery approach can benefit the occluded vehicle tracking in real-world videos.
****

FrameNet: Learning Local Canonical Frames of 3D Surfaces From a Single RGB Image
Jingwei Huang, Yichao Zhou, Thomas Funkhouser, Leonidas J. Guibas; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8638-8647
In this work, we introduce the novel problem of identifying dense canonical 3D coordinate frames from a single RGB image. We observe that each pixel in an image corresponds to a surface in the underlying 3D geometry, where a canonical frame can be identified as represented by three orthogonal axes, one along its normal direction and two in its tangent plane. We propose an algorithm to predict these axes from RGB. Our first insight is that canonical frames computed automatically with recently introduced direction field synthesis methods can provide training data for the task. Our second insight is that networks designed for surface normal prediction provide better results when trained jointly to predict canonical frames, and even better when trained to also predict 2D projections of canonical frames. We conjecture this is because projections of canonical tangent direct

ions often align with local gradients in images, and because those directions ar e tightly linked to 3Dcanonical frames through projective geometry and orthogona lity constraints. In our experiments, we find that our method predicts 3D canoni cal frames that can be used in applications ranging from surface normal estimati on, feature matching, and augmented reality.
****