IDA-3D: Instance-Depth-Aware 3D Object Detection From Stereo Vision for Autonomous Driving
Wanli Peng, Hao Pan, He Liu, Yi Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13015-13024
3D object detection is an important scene understanding task in autonomous driving and virtual reality. Approaches based on LiDAR technology have high performance, but LiDAR is expensive. Considering more general scenes, where there is no LiDAR data in the 3D datasets, we propose a 3D object detection approach from stereo vision which does not rely on LiDAR data either as input or as supervision in training, but solely takes RGB images with corresponding annotated 3D bounding boxes as training data. As depth estimation of object is the key factor affecting the performance of 3D object detection, we introduce an Instance-DepthAware (IDA) module which accurately predicts the depth of the 3D bounding box's center by instance-depth awareness, disparity adaptation and matching cost reweighting. Moreover, our model is an end-to-end learning framework which does not require multiple stages or postprocessing algorithm. We provide detailed experiments on KITTI benchmark and achieve impressive improvements compared with the existing image-based methods. Our code is available at https://github.com/swords123/IDA-3D.
****
FroDO: From Detections to 3D Objects
Martin Runz, Kejie Li, Meng Tang, Lingni Ma, Chen Kong, Tanner Schmidt, Ian Reid, Lourdes Agapito, Julian Straub, Steven Lovegrove, Richard Newcombe; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14720-14729
Object-oriented maps are important for scene understanding since they jointly capture geometry and semantics, allow individual instantiation and meaningful reasoning about objects. We introduce FroDO, a method for accurate 3D reconstruction of object instances from RGB video that infers their location, pose and shape in a coarse to fine manner. Key to FroDO is to embed object shapes in a novel learnt shape space that allows seamless switching between sparse point cloud and dense DeepSDF decoding. Given an input sequence of localized RGB frames, FroDO first aggregates 2D detections to instantiate a 3D bounding box per object. A shape code is regressed using an encoder network before optimizing shape and pose further under the learnt shape priors using sparse or dense shape representations. The optimization uses multi-view geometric, photometric and silhouette losses. We evaluate on real-world datasets, including Pix3D, Redwood-OS, and ScanNet, for single-view, multi-view, and multi-object reconstruction.
****
KeypointNet: A Large-Scale 3D Keypoint Dataset Aggregated From Numerous Human Annotations
Yang You, Yujing Lou, Chengkun Li, Zhoujun Cheng, Liangwei Li, Lizhuang Ma, Cewu Lu, Weiming Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13647-13656
Detecting 3D objects keypoints is ofgreat interest to the areas of both graphics and computer vision. There have been several 2D and 3D keypoint datasets aiming to address this problem in a data-driven way. These datasets, however, either lack scalability or bring ambiguity to the definition of keypoints. Therefore, we present KeypointNet: the first large-scale and diverse 3D keypoint dataset that contains 83,231 keypoints and 8,329 3D models from 16 object categories, by leveraging numerous human annotations. To handle the inconsistency between annotations from different people, we propose a novel method to aggregate these keypoints automatically, through minimization of a fidelity loss. Finally, ten state-of-the-art methods are benchmarked on our proposed dataset.
****
Bridging the Gap Between Anchor-Based and Anchor-Free Detection via Adaptive Training Sample Selection
Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, Stan Z. Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9759-9768

Object detection has been dominated by anchor-based detectors for several years. Recently, anchor-free detectors have become popular due to the proposal of FPN and Focal Loss. In this paper, we first point out that the essential difference between anchor-based and anchor-free detection is actually how to define positive and negative training samples, which leads to the performance gap between them. If they adopt the same definition of positive and negative samples during training, there is no obvious difference in the final performance, no matter regressing from a box or a point. This shows that how to select positive and negative training samples is important for current object detectors. Then, we propose an Adaptive Training Sample Selection (ATSS) to automatically select positive and negative samples according to statistical characteristics of object. It significantly improves the performance of anchor-based and anchor-free detectors and bridges the gap between them. Finally, we discuss the necessity of tiling multiple anchors per location on the image to detect objects. Extensive experiments conducted on MS COCO support our aforementioned analysis and conclusions. With the newly introduced ATSS, we improve state-of-the-art detectors by a large margin to 50.7% AP without introducing any overhead. The code is available at https://github.com/sfzhang15/ATSS.

****

Structure Aware Single-Stage 3D Object Detection From Point Cloud

Chenhang He, Hui Zeng, Jianqiang Huang, Xian-Sheng Hua, Lei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11873-11882

3D object detection from point cloud data plays an essential role in autonomous driving. Current single-stage detectors are efficient by progressively downscaling the 3D point clouds in a fully convolutional manner. However, the downscaled features inevitably lose spatial information and cannot make full use of the structure information of 3D point cloud, degrading their localization precision. In this work, we propose to improve the localization precision of single-stage detectors by explicitly leveraging the structure information of 3D point cloud. Specifically, we design an auxiliary network which converts the convolutional features in the backbone network back to point-level representations. The auxiliary network is jointly optimized, by two point-level supervisions, to guide the convolutional features in the backbone network to be aware of the object structure. The auxiliary network can be detached after training and therefore introduces no extra computation in the inference stage. Besides, considering that single-stage detectors suffer from the discordance between the predicted bounding boxes and corresponding classification confidences, we develop an efficient part-sensitive warping operation to align the confidences to the predicted bounding boxes. Our proposed detector ranks at the top of KITTI 3D/BEV detection leaderboards and runs at 25 FPS for inference.

****

MINA: Convex Mixed-Integer Programming for Non-Rigid Shape Alignment

Florian Bernard, Zeeshan Khan Suri, Christian Theobalt; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13826-13835

We present a convex mixed-integer programming formulation for non-rigid shape matching. To this end, we propose a novel shape deformation model based on an efficient low-dimensional discrete model, so that finding a globally optimal solution is tractable in (most) practical cases. Our approach combines several favourable properties, namely it is independent of the initialisation, it is much more efficient to solve to global optimality compared to analogous quadratic assignment problem formulations, and it is highly flexible in terms of the variants of matching problems it can handle. Experimentally we demonstrate that our approach outperforms existing methods for sparse shape matching, that it can be used for initialising dense shape matching methods, and we showcase its flexibility on several examples.

****

Enhanced Transport Distance for Unsupervised Domain Adaptation

Mengxue Li, Yi-Ming Zhai, You-Wei Luo, Peng-Fei Ge, Chuan-Xian Ren; Proceedi

ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13936-13944
Unsupervised domain adaptation (UDA) is a representative problem in transfer learning, which aims to improve the classification performance on an unlabeled target domain by exploiting discriminant information from a labeled source domain. The optimal transport model has been used for UDA in the perspective of distribution matching. However, the transport distance cannot reflect the discriminant information from either domain knowledge or category prior. In this work, we propose an enhanced transport distance (ETD) for UDA. This method builds an attention-aware transport distance, which can be viewed as the prediction feedback of the iteratively learned classifier, to measure the domain discrepancy. Further, the Kantorovich potential variable is re-parameterized by deep neural networks to learn the distribution in the latent space. The entropy-based regularization is developed to explore the intrinsic structure of the target domain. The proposed method is optimized alternately in an end-to-end manner. Extensive experiments are conducted on four benchmark datasets to demonstrate the SOTA performance of ETD.
****

A Local-to-Global Approach to Multi-Modal Movie Scene Segmentation
Anyi Rao, Linning Xu, Yu Xiong, Guodong Xu, Qingqiu Huang, Bolei Zhou, Dahua Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10146-10155
Scene, as the crucial unit of storytelling in movies, contains complex activities of actors and their interactions in a physical environment. Identifying the composition of scenes serves as a critical step towards semantic understanding of movies. This is very challenging - compared to the videos studied in conventional vision problems, e.g. action recognition, as scenes in movies usually contain much richer temporal structures and more complex semantic information. Towards this goal, we scale up the scene segmentation task by building a large-scale video dataset MovieScenes, which contains 21K annotated scene segments from 150 movies. We further propose a local-to-global scene segmentation framework, which integrates multi-modal information across three levels, i.e. clip, segment, and movie. This framework is able to distill complex semantics from hierarchical temporal structures over a long movie, providing top-down guidance for scene segmentation. Our experiments show that the proposed network is able to segment a movie into scenes with high accuracy, consistently outperforming previous methods. We also found that pretraining on our MovieScenes can bring significant improvements to the existing approaches.
****

TA-Student VQA: Multi-Agents Training by Self-Questioning
Peixi Xiong, Ying Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10065-10075
There are two main challenges in Visual Question Answering (VQA). The first one is that each model obtains its strengths and shortcomings when applied to several questions; what is more, the "ceiling effect" for specific questions is difficult to overcome with simple consecutive training. The second challenge is that even the state-of-the-art dataset is of large scale, questions targeted at a single image are off in format and lack diversity in content. We introduce our self-questioning model with multi-agent training: TA-student VQA. This framework differs from standard VQA algorithms by involving question-generating mechanisms and collaborative learning questions between question-answering agents. Thus, TA-student VQA overcomes the limitation of the content diversity and format variation of questions and improves the overall performance of multiple question-answering agents. We evaluate our model on VQA-v2, which outperforms algorithms without such mechanisms. In addition, TA-student VQA achieves a greater model capacity, allowing it to answer more generated questions in addition to those in the annotated datasets.
****

Deep Structure-Revealed Network for Texture Recognition
Wei Zhai, Yang Cao, Zheng-Jun Zha, HaiYong Xie, Feng Wu; Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11010-11019

Texture recognition is a challenging visual task since various primitives along with their arrangements can be recognized from a same texture image when perceiving with different contexts. Some recent work building on CNNs exploits orderless aggregating to provide invariance to spatial arrangements. However, these methods ignore the inherent structural property of textures, which is a critical cue for distinguishing and describing texture images in the wild. To address this problem, we propose a novel Deep Structure-Revealed Network (DSR-Net) that leverages spatial dependency among the captured primitives as structural representation for texture recognition. Specifically, a primitive capturing module (PCM) is devised to generate multiple primitives from eight directional spatial contexts, in which deep features are firstly extracted under the constrains of direction map and then encoded based on the similarities of neighborhood. Next, these primitives are associated with a dependence learning module (DLM) to generate structural representation, in which a two-way collaborative relationship strategy is introduced to perceive the spatial dependencies among multiple primitives. At last, the structure-revealed texture representations are integrated with spatial ordered information to achieve real-world texture recognition. Evaluation on the five most challenging texture recognition datasets has demonstrated the superiority of the proposed model against state-of-the-art methods. The structure-revealed performances of DSR-Net are further verified on some extensive experiments, including fine-grained classification and semantic segmentation.

****

Dynamic Traffic Modeling From Overhead Imagery

Scott Workman, Nathan Jacobs; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12315-12324

Our goal is to use overhead imagery to understand patterns in traffic flow, for instance answering questions such as how fast could you traverse Times Square at 3am on a Sunday. A traditional approach for solving this problem would be to model the speed of each road segment as a function of time. However, this strategy is limited in that a significant amount of data must first be collected before a model can be used and it fails to generalize to new areas. Instead, we propose an automatic approach for generating dynamic maps of traffic speeds using convolutional neural networks. Our method operates on overhead imagery, is conditioned on location and time, and outputs a local motion model that captures likely directions of travel and corresponding travel speeds. To train our model, we take advantage of historical traffic data collected from New York City. Experimental results demonstrate that our method can be applied to generate accurate city-scale traffic models.

****

In Defense of Grid Features for Visual Question Answering

Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, Xinlei Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10267-10276

Popularized as `bottom-up' attention, bounding box (or region) based visual features have recently surpassed vanilla grid-based convolutional features as the de facto standard for vision and language tasks like visual question answering (VQA). However, it is not clear whether the advantages of regions (e.g. better localization) are the key reasons for the success of bottom-up attention. In this paper, we revisit grid features for VQA, and find they can work surprisingly well -- running more than an order of magnitude faster with the same accuracy (e.g. if pre-trained in a similar fashion). Through extensive experiments, we verify that this observation holds true across different VQA models (reporting a state-of-the-art accuracy on VQA 2.0 test-std, 72.71), datasets, and generalizes well to other tasks like image captioning. As grid features make the model design and training process much simpler, this enables us to train them end-to-end and also use a more flexible network design. We learn VQA models end-to-end, from pixels directly to answers, and show that strong performance is achievable without using any region annotations in pre-training. We hope our findings help further impr

ove the scientific understanding and the practical application of VQA. Code and features will be made available.
****

Rethinking the Route Towards Weakly Supervised Object Localization
Chen-Lin Zhang, Yun-Hao Cao, Jianxin Wu; Proceedings of the IEEE/CVF Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13460-13469
Weakly supervised object localization (WSOL) aims to localize objects with only image-level labels. Previous methods often try to utilize feature maps and class ification weights to localize objects using image level annotations indirectly. In this paper, we demonstrate that weakly supervised object localization should be divided into two parts: class-agnostic object localization and object classif ication. For class-agnostic object localization, we should use class-agnostic me thods to generate noisy pseudo annotations and then perform bounding box regress ion on them without class labels. We propose the pseudo supervised object locali zation (PSOL) method as a new way to solve WSOL. Our PSOL models have good trans ferability across different datasets without fine-tuning. With generated pseudo bounding boxes, we achieve 58.00% localization accuracy on ImageNet and 74.74% l ocalization accuracy on CUB-200, which have a large edge over previous models.
****

Weakly Supervised Fine-Grained Image Classification via Guassian Mixture Model O riented Discriminative Learning
Zhihui Wang, Shijie Wang, Shuhui Yang, Haojie Li, Jianjun Li, Zezhou Li; Pr oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9749-9758
Existing weakly supervised fine-grained image recognition (WFGIR) methods usuall y pick out the discriminative regions from the high-level feature maps directly. We discover that due to the operation of stacking local receptive filed, Convol utional Neural Network causes the discriminative region diffusion in high-level feature maps, which leads to inaccurate discriminative region localization. In t his paper, we propose an end-to-end Discriminative Feature-oriented Gaussian Mix ture Model (DF-GMM), to address the problem of discriminative region diffusion a nd find better fine-grained details. Specifically, DF-GMM consists of 1) a low-r ank representation mechanism (LRM), which learns a set of low-rank discriminativ e bases by Gaussian Mixture Model (GMM) in high-level semantic feature maps to i mprove discriminative ability of feature representation, 2) a low-rank represent ation reorganization mechanism (LR ^2 M) which resumes the space information cor responding to low-rank discriminative bases to reconstruct the low-rank feature maps. It alleviates the discriminative region diffusion problem and locate discr iminative regions more precisely. Extensive experiments verify that DF-GMM yield s the best performance under the same settings with the most competitive approac hes, in CUB-Bird, Stanford-Cars datasets, and FGVC Aircraft.
****

Violin: A Large-Scale Dataset for Video-and-Language Inference
Jingzhou Liu, Wenhu Chen, Yu Cheng, Zhe Gan, Licheng Yu, Yiming Yang, Jing jing Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10900-10910
We introduce a new task, Video-and-Language Inference, for joint multimodal unde rstanding of video and text. Given a video clip with aligned subtitles as premis e, paired with a natural language hypothesis based on the video content, a model needs to infer whether the hypothesis is entailed or contradicted by the given video clip. A new large-scale dataset, named Violin (VIdeO-and-Language INferenc e), is introduced for this task, which consists of 95,322 video-hypothesis pairs from 15,887 video clips, spanning over 582 hours of video. These video clips co ntain rich content with diverse temporal dynamics, event shifts, and people inte ractions, collected from two sources: (i) popular TV shows, and (ii) movie clips from YouTube channels. In order to address our new multimodal inference task, a model is required to possess sophisticated reasoning skills, from surface-level grounding (e.g., identifying objects and characters in the video) to in-depth c ommonsense reasoning (e.g., inferring causal relations of events in the video). We present a detailed analysis of the dataset and an extensive evaluation over m

any strong baselines, providing valuable insights on the challenges of this new task.
****
Local Context Normalization: Revisiting Local Normalization
Anthony Ortiz, Caleb Robinson, Dan Morris, Olac Fuentes, Christopher Kiekintveld, Md Mahmudulla Hassan, Nebojsa Jojic; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11276-11285
Normalization layers have been shown to improve convergence in deep neural networks, and even add useful inductive biases. In many vision applications the local spatial context of the features is important, but most common normalization schemes including Group Normalization (GN), Instance Normalization (IN), and Layer Normalization (LN) normalize over the entire spatial dimension of a feature. This can wash out important signals and degrade performance. For example, in applications that use satellite imagery, input images can be arbitrarily large; consequently, it is nonsensical to normalize over the entire area. Positional Normalization (PN), on the other hand, only normalizes over a single spatial position at a time. A natural compromise is to normalize features by local context, while also taking into account group level information. In this pa- per, we propose Local Context Normalization (LCN): a normalization layer where every feature is normalized based on a window around it and the filters in its group. We propose an algorithmic solution to make LCN efficient for arbitrary window sizes, even if every point in the image has a unique window. LCN outperforms its Batch Normalization (BN), GN, IN, and LN counterparts for object detection, semantic segmentation, and instance segmentation applications in several benchmark datasets, while keeping performance in- dependent of the batch size and facilitating transfer learning.
****
Tangent Images for Mitigating Spherical Distortion
Marc Eder, Mykhailo Shvets, John Lim, Jan-Michael Frahm; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12426-12434
In this work, we propose "tangent images," a spherical image representation that facilitates transferable and scalable 360 degree computer vision. Inspired by techniques in cartography and computer graphics, we render a spherical image to a set of distortion-mitigated, locally-planar image grids tangent to a subdivided icosahedron. By varying the resolution of these grids independently of the subdivision level, we can effectively represent high resolution spherical images while still benefiting from the low-distortion icosahedral spherical approximation. We show that training standard convolutional neural networks on tangent images compares favorably to the many specialized spherical convolutional kernels that have been developed, while also scaling efficiently to handle significantly higher spherical resolutions. Furthermore, because our approach does not require specialized kernels, we show that we can transfer networks trained on perspective images to spherical data without fine-tuning and with limited performance drop-off. Finally, we demonstrate that tangent images can be used to improve the quality of sparse feature detection on spherical images, illustrating its usefulness for traditional computer vision tasks like structure-from-motion and SLAM.
****
Normalized and Geometry-Aware Self-Attention Network for Image Captioning
Longteng Guo, Jing Liu, Xinxin Zhu, Peng Yao, Shichen Lu, Hanqing Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10327-10336
Self-attention (SA) network has shown profound value in image captioning. In this paper, we improve SA from two aspects to promote the performance of image captioning. First, we propose Normalized Self-Attention (NSA), a reparameterization of SA that brings the benefits of normalization inside SA. While normalization is previously only applied outside SA, we introduce a novel normalization method and demonstrate that it is both possible and beneficial to perform it on the hidden activations inside SA. Second, to compensate for the major limit of Transformer that it fails to model the geometry structure of the input objects, we propo

se a class of Geometry-aware Self-Attention (GSA) that extends SA to explicitly and efficiently consider the relative geometry relations between the objects in the image. To construct our image captioning model, we combine the two modules a nd apply it to the vanilla self-attention network. We extensively evaluate our p roposals on MS-COCO image captioning dataset and superior results are achieved w hen comparing to state-of-the-art approaches. Further experiments on three chall enging tasks, i.e. video captioning, machine translation, and visual question an swering, show the generality of our methods.
****

Deepstrip: High-Resolution Boundary Refinement
Peng Zhou, Brian Price, Scott Cohen, Gregg Wilensky, Larry S. Davis; Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10558-10567
In this paper, we target refining the boundaries in high resolution images given low resolution masks. For memory and computation efficiency, we propose to conv ert the regions of interest into strip images and compute a boundary prediction in the strip domain. To detect the target boundary, we present a framework with two prediction layers. First, all potential boundaries are predicted as an initi al prediction and then a selection layer is used to pick the target boundary and smooth the result. To encourage accurate prediction, a loss which measures the boundary distance in strip domain is introduced. In addition, we enforce a match ing consistency and C0 continuity regularization to the network to reduce false alarms. Extensive experiments on both public and a newly created high resolution dataset strongly validate our approach.
****

xMUDA: Cross-Modal Unsupervised Domain Adaptation for 3D Semantic Segmentation
Maximilian Jaritz, Tuan-Hung Vu, Raoul de Charette, Emilie Wirbel, Patrick P erez; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Reco gnition (CVPR), 2020, pp. 12605-12614
Unsupervised Domain Adaptation (UDA) is crucial to tackle the lack of annotation s in a new domain. There are many multi-modal datasets, but most UDA approaches are uni-modal. In this work, we explore how to learn from multi-modality and pro pose cross-modal UDA (xMUDA) where we assume the presence of 2D images and 3D po int clouds for 3D semantic segmentation. This is challenging as the two input sp aces are heterogeneous and can be impacted differently by domain shift. In xMUDA , modalities learn from each other through mutual mimicking, disentangled from t he segmentation objective, to prevent the stronger modality from adopting false predictions from the weaker one. We evaluate on new UDA scenarios including day- to-night, country-to-country and dataset-to-dataset, leveraging recent autonomou s driving datasets. xMUDA brings large improvements over uni-modal UDA on all te sted scenarios, and is complementary to state-of-the-art UDA techniques. Code is available at https://github.com/valeoai/xmuda.
****

Seeing without Looking: Contextual Rescoring of Object Detections for AP Maximiz ation
Lourenco V. Pato, Renato Negrinho, Pedro M. Q. Aguiar; Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 146 10-14618
The majority of current object detectors lack context: class predictions are mad e independently from other detections. We propose to incorporate context in obje ct detection by post-processing the output of an arbitrary detector to rescore t he confidences of its detections. Rescoring is done by conditioning on contextua l information from the entire set of detections: their confidences, predicted cl asses, and positions. We show that AP can be improved by simply reassigning the detection confidence values such that true positives that survive longer (i.e., those with the correct class and large IoU) are scored higher than false positiv es or detections with small IoU. In this setting, we use a bidirectional RNN wit h attention for contextual rescoring and introduce a training target that uses t he IoU with ground truth to maximize AP for the given set of detections. The fac t that our approach does not require access to visual features makes it computat

ionally inexpensive and agnostic to the detection architecture. In spite of this simplicity, our model consistently improves AP over strong pre-trained baselines (Cascade R-CNN and Faster R-CNN with several backbones), particularly by reducing the confidence of duplicate detections (a learned form of non-maximum suppression) and removing out-of-context objects by conditioning on the confidences, classes, positions, and sizes of the co-occurrent detections. Code is available at https://github.com/LourencoVazPato/seeing-without-looking/
****

Learning Geocentric Object Pose in Oblique Monocular Images
Gordon Christie, Rodrigo Rene Rai Munoz Abujder, Kevin Foster, Shea Hagstrom, Gregory D. Hager, Myron Z. Brown; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14512-14520
An object's geocentric pose, defined as the height above ground and orientation with respect to gravity, is a powerful representation of real-world structure for object detection, segmentation, and localization tasks using RGBD images. For close-range vision tasks, height and orientation have been derived directly from stereo-computed depth and more recently from monocular depth predicted by deep networks. For long-range vision tasks such as Earth observation, depth cannot be reliably estimated with monocular images. Inspired by recent work in monocular height above ground prediction and optical flow prediction from static images, we develop an encoding of geocentric pose to address this challenge and train a deep network to compute the representation densely, supervised by publicly available airborne lidar. We exploit these attributes to rectify oblique images and remove observed object parallax to dramatically improve the accuracy of localization and to enable accurate alignment of multiple images taken from very different oblique viewpoints. We demonstrate the value of our approach by extending two large-scale public datasets for semantic segmentation in oblique satellite images. All of our data and code are publicly available.
****

DPGN: Distribution Propagation Graph Network for Few-Shot Learning
Ling Yang, Liangliang Li, Zilun Zhang, Xinyu Zhou, Erjin Zhou, Yu Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13390-13399
Most graph-network-based meta-learning approaches model instance-level relation of examples. We extend this idea further to explicitly model the distribution-level relation of one example to all other examples in a 1-vs-N manner. We propose a novel approach named distribution propagation graph network (DPGN) for few-shot learning. It conveys both the distribution-level relations and instance-level relations in each few-shot learning task. To combine the distribution-level relations and instance-level relations for all examples, we construct a dual complete graph network which consists of a point graph and a distribution graph with each node standing for an example. Equipped with dual graph architecture, DPGN propagates label information from labeled examples to unlabeled examples within several update generations. In extensive experiments on few-shot learning benchmarks, DPGN outperforms state-of-the-art results by a large margin in 5% 12% under supervised setting and 7% 13% under semi-supervised setting. Code will be released.
****

Cross-Domain Document Object Detection: Benchmark Suite and Method
Kai Li, Curtis Wigington, Chris Tensmeyer, Handong Zhao, Nikolaos Barmpalios, Vlad I. Morariu, Varun Manjunatha, Tong Sun, Yun Fu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12915-12924
Decomposing images of document pages into high-level semantic regions (e.g., figures, tables, paragraphs), document object detection (DOD) is fundamental for downstream tasks like intelligent document editing and understanding. DOD remains a challenging problem as document objects vary significantly in layout, size, aspect ratio, texture, etc. An additional challenge arises in practice because large labeled training datasets are only available for domains that differ from the target domain. We investigate cross-domain DOD, where the goal is to learn a de

tector for the target domain using labeled data from the source domain and only unlabeled data from the target domain. Documents from the two domains may vary significantly in layout, language, and genre. We establish a benchmark suite consisting of different types of PDF document datasets that can be utilized for cross-domain DOD model training and evaluation. For each dataset, we provide the page images, bounding box annotations, PDF files, and the rendering layers extracted from the PDF files. Moreover, we propose a novel cross-domain DOD model which builds upon the standard detection model and addresses domain shifts by incorporating three novel alignment modules: Feature Pyramid Alignment (FPA) module, Region Alignment (RA) module and Rendering Layer alignment (RLA) module. Extensive experiments on the benchmark suite substantiate the efficacy of the three proposed modules and the proposed method significantly outperforms the baseline methods. The project page is at https://github.com/kailigo/cddod.
****

On Translation Invariance in CNNs: Convolutional Layers Can Exploit Absolute Spatial Location
Osman Semih Kayhan, Jan C. van Gemert; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14274-14285
In this paper we challenge the common assumption that convolutional layers in modern CNNs are translation invariant. We show that CNNs can and will exploit the absolute spatial location by learning filters that respond exclusively to particular absolute locations by exploiting image boundary effects. Because modern CNNs filters have a huge receptive field, these boundary effects operate even far from the image boundary, allowing the network to exploit absolute spatial location all over the image. We give a simple solution to remove spatial location encoding which improves translation invariance and thus gives a stronger visual inductive bias which particularly benefits small data sets. We broadly demonstrate these benefits on several architectures and various applications such as image classification, patch matching, and two video classification datasets.
****

Classifying, Segmenting, and Tracking Object Instances in Video with Mask Propagation
Gedas Bertasius, Lorenzo Torresani; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9739-9748
We introduce a method for simultaneously classifying, segmenting and tracking object instances in a video sequence. Our method, named MaskProp, adapts the popular Mask R-CNN to video by adding a mask propagation branch that propagates frame-level object instance masks from each video frame to all the other frames in a video clip. This allows our system to predict clip-level instance tracks with respect to the object instances segmented in the middle frame of the clip. Clip-level instance tracks generated densely for each frame in the sequence are finally aggregated to produce video-level object instance segmentation and classification. Our experiments demonstrate that our clip-level instance segmentation makes our approach robust to motion blur and object occlusions in video. MaskProp achieves the best reported accuracy on the YouTube-VIS dataset, outperforming the ICCV 2019 video instance segmentation challenge winner despite being much simpler and using orders of magnitude less labeled data (1.3M vs 1B images and 860K vs 14M bounding boxes). The project page is at: https://gberta.github.io/maskprop/.
****

WaveletStereo: Learning Wavelet Coefficients of Disparity Map in Stereo Matching
Menglong Yang, Fangrui Wu, Wei Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12885-12894
Some stereo matching algorithms based on deep learning have been proposed and achieved state-of-the-art performances since some public large-scale datasets were put online. However, the disparity in smooth regions and detailed regions is still difficult to accurately estimate simultaneously. This paper proposes a novel stereo matching method called WaveletStereo, which learns the wavelet coefficients of the disparity rather than the disparity itself. The WaveletStereo consists of several sub-modules, where the low-frequency sub-module generates the low-frequency wavelet coefficients, which aims at learning global context information

and well handling the low-frequency regions such as textureless surfaces, and the others focus on the details. In addition, a densely connected atrous spatial pyramid block is introduced for better learning the multi-scale image features. Experimental results show the effectiveness of the proposed method, which achieves state-of-the-art performance on the large-scale test dataset Scene Flow.
****

Connect-and-Slice: An Hybrid Approach for Reconstructing 3D Objects
Hao Fang, Florent Lafarge; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13490-13498
Converting point clouds generated by Laser scanning, multiview stereo imagery or depth cameras into compact polygon meshes is a challenging problem in vision. Existing methods are either robust to imperfect data or scalable, but rarely both. In this paper, we address this issue with an hybrid method that successively connects and slices planes detected from 3D data. The core idea consists in constructing an efficient and compact partitioning data structure. The later is i) spatially-adaptive in the sense that a plane slices a restricted number of relevant planes only, and ii) composed of components with different structural meaning resulting from a preliminary analysis of the plane connectivity. Our experiments on a variety of objects and sensors show the versatility of our approach as well as its competitiveness with respect to existing methods.
****

Learning Instance Occlusion for Panoptic Segmentation
Justin Lazarow, Kwonjoon Lee, Kunyu Shi, Zhuowen Tu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10720-10729
Panoptic segmentation requires segments of both "things" (countable object instances) and "stuff" (uncountable and amorphous regions) within a single output. A common approach involves the fusion of instance segmentation (for "things") and semantic segmentation (for "stuff") into a non-overlapping placement of segments, and resolves overlaps. However, instance ordering with detection confidence does not correlate well with natural occlusion relationship. To resolve this issue, we propose a branch that is tasked with modeling how two instance masks should overlap one another as a binary relation. Our method, named OCFusion, is lightweight but particularly effective in the instance fusion process. OCFusion is trained with the ground truth relation derived automatically from the existing dataset annotations. We obtain state-of-the-art results on COCO and show competitive results on the Cityscapes panoptic segmentation benchmark.
****

Graph Embedded Pose Clustering for Anomaly Detection
Amir Markovitz, Gilad Sharir, Itamar Friedman, Lihi Zelnik-Manor, Shai Avidan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10539-10547
We propose a new method for anomaly detection of human actions. Our method works directly on human pose graphs that can be computed from an input video sequence. This makes the analysis independent of nuisance parameters such as viewpoint or illumination. We map these graphs to a latent space and cluster them. Each action is then represented by its soft-assignment to each of the clusters. This gives a kind of "bag of words" representation to the data, where every action is represented by its similarity to a group of base action-words. Then, we use a Dirichlet process based mixture, that is useful for handling proportional data such as our soft-assignment vectors, to determine if an action is normal or not. We evaluate our method on two types of data sets. The first is a fine-grained anomaly detection data set (e.g. ShanghaiTech) where we wish to detect unusual variations of some action. The second is a coarse-grained anomaly detection data set (e.g., a Kinetics-based data set) where few actions are considered normal, and every other action should be considered abnormal. Extensive experiments on the benchmarks show that our method1performs considerably better than other state of the art methods.
****

Graph Structured Network for Image-Text Matching

Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, Yongdong Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10921-10930

Image-text matching has received growing interest since it bridges vision and language. The key challenge lies in how to learn correspondence between image and text. Existing works learn coarse correspondence based on object co-occurrence statistics, while failing to learn fine-grained phrase correspondence. In this paper, we present a novel Graph Structured Matching Network (GSMN) to learn fine-grained correspondence. The GSMN explicitly models object, relation and attribute as a structured phrase, which not only allows to learn correspondence of object, relation and attribute separately, but also benefits to learn fine-grained correspondence of structured phrase. This is achieved by node-level matching and structure-level matching. The node-level matching associates each node with its relevant nodes from another modality, where the node can be object, relation or attribute. The associated nodes then jointly infer fine-grained correspondence by fusing neighborhood associations at structure-level matching. Comprehensive experiments show that GSMN outperforms state-of-the-art methods on benchmarks, with relative Recall@1 improvements of nearly 7% and 2% on Flickr30K and MSCOCO, respectively. Code will be released at: https://github.com/CrossmodalGroup/GSMN.

****

BFBox: Searching Face-Appropriate Backbone and Feature Pyramid Network for Face Detector

Yang Liu, Xu Tang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13568-13577

Popular backbones designed on image classification have demonstrated their considerable compatibility on the task of general object detection. However, the same phenomenon does not appear on the face detection. This is largely due to the average scale of ground-truth in the WiderFace dataset is far smaller than that of generic objects in theCOCO one. To resolve this, the success of Neural Archi-tecture Search (NAS) inspires us to search face-appropriate backbone and featrue pyramid network (FPN) architecture.Firstly, we design the search space for backbone and FPN by comparing performance of feature maps with different backbones and excellent FPN architectures on the face detection. Second, we propose a FPN-attention module to joint search the architecture of backbone and FPN. Finally,we conduct comprehensive experiments on popular bench-marks, including Wider Face, FDDB, AFW and PASCALFace, display the superiority of our proposed method.

****

End-to-End Adversarial-Attention Network for Multi-Modal Clustering

Runwu Zhou, Yi-Dong Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14619-14628

Multi-modal clustering aims to cluster data into different groups by exploring complementary information from multiple modalities or views. Little work learns the deep fused representations and simutaneously discovers the cluster structure with a discriminative loss. In this paper, we present an End-to-end Adversarial-attention network for Multi-modal Clustering (EAMC), where adversarial learning and attention mechanism are leveraged to align the latent feature distributions and quantify the importance of modalities respectively. To benefit from the joint training, we introducea divergence-based clustering objective that not only encourages the separation and compactness of the clusters but also enjoy a clear cluster structure by embedding the simplex geometry of the output space into the loss. The proposed network consists of modality-specific feature learning, modality fusion and cluster assignment three modules. It can be trained from scratch with batch-mode based optimization and avoid an autoencoder pretraining stage. Comprehensive experiments conducted on five real-world datasets show the superiority and effectiveness of the proposed clustering method.

****

What You See is What You Get: Exploiting Visibility for 3D Object Detection

Peiyun Hu, Jason Ziglar, David Held, Deva Ramanan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11001-11009

Recent advances in 3D sensing have created unique challenges for computer vision . One fundamental challenge is finding a good representation for 3D sensor data. Most popular representations (such as PointNet) are proposed in the context of processing truly 3D data (e.g. points sampled from mesh models), ignoring the fa ct that 3D sensored data such as a LiDAR sweep is in fact 2.5D. We argue that re presenting 2.5D data as collections of (x,y,z) points fundamentally destroys hid den information about freespace. In this paper, we demonstrate such knowledge ca n be efficiently recovered through 3D raycasting and readily incorporated into b atch-based gradient learning. We describe a simple approach to augmenting voxel-based networks with visibility: we add a voxelized visibility map as an addition al input stream. In addition, we show that visibility can be combined with two c rucial modifications common to state-of-the-art 3D detectors: synthetic data aug mentation of virtual objects and temporal aggregation of LiDAR sweeps over multi ple time frames. On the NuScenes 3D detection benchmark, we show that, by adding an additional stream for visibility input, we can significantly improve the ove rall detection accuracy of a state-of-the-art 3D detector.
****

Visual Grounding in Video for Unsupervised Word Translation
Gunnar A. Sigurdsson, Jean-Baptiste Alayrac, Aida Nematzadeh, Lucas Smaira, Mateusz Malinowski, Joao Carreira, Phil Blunsom, Andrew Zisserman; Proceeding s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10850-10859
There are thousands of actively spoken languages on Earth, but a single visual w orld. Grounding in this visual world has the potential to bridge the gap between all these languages. Our goal is to use visual grounding to improve unsupervise d word mapping between languages. The key idea is to establish a common visual r epresentation between two languages by learning embeddings from unpaired instruc tional videos narrated in the native language. Given this shared embedding we de monstrate that (i) we can map words between the languages, particularly the 'vis ual' words; (ii) that the shared embedding provides a good initialization for ex isting unsupervised text-based word translation techniques, forming the basis fo r our proposed hybrid visual-text mapping algorithm, MUVE; and (iii) our approac h achieves superior performance by addressing the shortcomings of text-based met hods -- it is more robust, handles datasets with less commonality, and is applic able to low-resource languages. We apply these methods to translate words from E nglish to French, Korean, and Japanese -- all without any parallel corpora and s imply by watching many videos of people speaking while doing things.
****

Learning Individual Speaking Styles for Accurate Lip to Speech Synthesis
K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, C.V. Jawahar; Procee dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP R), 2020, pp. 13796-13805
Humans involuntarily tend to infer parts of the conversation from lip movements when the speech is absent or corrupted by external noise. In this work, we explo re the task of lip to speech synthesis, i.e., learning to generate natural speec h given only the lip movements of a speaker. Acknowledging the importance of con textual and speaker-specific cues for accurate lip-reading, we take a different path from existing works. We focus on learning accurate lip sequences to speech mappings for individual speakers in unconstrained, large vocabulary settings. To this end, we collect and release a large-scale benchmark dataset, the first of its kind, specifically to train and evaluate the single-speaker lip to speech ta sk in natural settings. We propose a novel approach with key design choices to a chieve accurate, natural lip to speech synthesis in such unconstrained scenarios for the first time. Extensive evaluation using quantitative, qualitative metric s and human evaluation shows that our method is four times more intelligible tha n previous works in this space.
****

Adversarial Latent Autoencoders
Stanislav Pidhorskyi, Donald A. Adjeroh, Gianfranco Doretto; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, p

p. 14104-14113
Autoencoder networks are unsupervised approaches aiming at combining generative and representational properties by learning simultaneously an encoder-generator map. Although studied extensively, the issues of whether they have the same generative power of GANs, or learn disentangled representations, have not been fully addressed. We introduce an autoencoder that tackles these issues jointly, which we call Adversarial Latent Autoencoder (ALAE). It is a general architecture that can leverage recent improvements on GAN training procedures. We designed two autoencoders: one based on a MLP encoder, and another based on a StyleGAN generator, which we call StyleALAE. We verify the disentanglement properties of both architectures. We show that StyleALAE can not only generate 1024x1024 face images with comparable quality of StyleGAN, but at the same resolution can also produce face reconstructions and manipulations based on real images. This makes ALAE the first autoencoder able to compare with, and go beyond the capabilities of a generator-only type of architecture.
****

Counterfactual Samples Synthesizing for Robust Visual Question Answering
Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, Yueting Zhuang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10800-10809
Despite Visual Question Answering (VQA) has realized impressive progress over the last few years, today's VQA models tend to capture superficial linguistic correlations in the train set and fail to generalize to the test set with different QA distributions. To reduce the language biases, several recent works introduce an auxiliary question-only model to regularize the training of targeted VQA model, and achieve dominating performance on VQA-CP. However, since the complexity of design, current methods are unable to equip the ensemble-based models with two indispensable characteristics of an ideal VQA model: 1) visual-explainable: the model should rely on the right visual regions when making decisions. 2) question-sensitive: the model should be sensitive to the linguistic variations in question. To this end, we propose a model-agnostic Counterfactual Samples Synthesizing (CSS) training scheme. The CSS generates numerous counterfactual training samples by masking critical objects in images or words in questions, and assigning different ground-truth answers. After training with the complementary samples (ie, the original and generated samples), the VQA models are forced to focus on all critical objects and words, which significantly improves both visual-explainable and question-sensitive abilities. In return, the performance of these models is further boosted. Extensive ablations have shown the effectiveness of CSS. Particularly, by building on top of the model LMH, we achieve a record-breaking performance of 58.95% on VQA-CP v2, with 6.5% gains.
****

Inter-Region Affinity Distillation for Road Marking Segmentation
Yuenan Hou, Zheng Ma, Chunxiao Liu, Tak-Wai Hui, Chen Change Loy; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12486-12495
We study the problem of distilling knowledge from a large deep teacher network to a much smaller student network for the task of road marking segmentation. In this work, we explore a novel knowledge distillation (KD) approach that can transfer 'knowledge' on scene structure more effectively from a teacher to a student model. Our method is known as Inter-Region Affinity KD (IntRA-KD). It decomposes a given road scene image into different regions and represents each region as a node in a graph. An inter-region affinity graph is then formed by establishing pairwise relationships between nodes based on their similarity in feature distribution. To learn structural knowledge from the teacher network, the student is required to match the graph generated by the teacher. The proposed method shows promising results on three large-scale road marking segmentation benchmarks, i.e., ApolloScape, CULane and LLAMAS, by taking various lightweight models as students and ResNet-101 as the teacher. IntRA-KD consistently brings higher performance gains on all lightweight models, compared to previous distillation methods. Our code is available at https://github.com/ cardwing/Codes-for-IntRA-KD.

****
Deformation-Aware Unpaired Image Translation for Pose Estimation on Laboratory Animals
Siyuan Li, Semih Gunel, Mirela Ostrek, Pavan Ramdya, Pascal Fua, Helge Rhodin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13158-13168
Our goal is to capture the pose of real animals using synthetic training examples, without using any manual supervision. Our focus is on neuroscience model organisms, to be able to study how neural circuits orchestrate behaviour. Human pose estimation attains remarkable accuracy when trained on real or simulated datasets consisting of millions of frames. However, for many applications simulated models are unrealistic and real training datasets with comprehensive annotations do not exist. We address this problem with a new sim2real domain transfer method. Our key contribution is the explicit and independent modeling of appearance, shape and pose in an unpaired image translation framework. Our model lets us train a pose estimator on the target domain by transferring readily available body keypoint locations from the source domain to generated target images. We compare our approach with existing domain transfer methods and demonstrate improved pose estimation accuracy on Drosophila melanogaster (fruit fly), Caenorhabditis elegans (worm) and Danio rerio (zebrafish), without requiring any manual annotation on the target domain and despite using simplistic off-the-shelf animal characters for simulation, or simple geometric shapes as models. Our new datasets, code and trained models will be published to support future computer vision and neuroscientific studies.
****
Few-Shot Pill Recognition
Suiyi Ling, Andreas Pastor, Jing Li, Zhaohui Che, Junle Wang, Jieun Kim, Patrick Le Callet; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9789-9798
Pill image recognition is vital for many personal/public health-care applications and should be robust to diverse unconstrained real-world conditions. Most existing pill recognition models are limited in tackling this challenging few-shot learning problem due to the insufficient instances per category. With limited training data, neural network-based models have limitations in discovering most discriminating features, or going deeper. Especially, existing models fail to handle the hard samples taken under less controlled imaging conditions. In this study, a new pill image database, namely CURE, is first developed with more varied imaging conditions and instances for each pill category. Secondly, a W2-net is proposed for better pill segmentation. Thirdly, a Multi-Stream (MS) deep network that captures task-related features along with a novel two-stage training methodology are proposed. Within the proposed framework, a Batch All strategy that considers all the samples is first employed for the sub-streams, and then a Batch Hard strategy that considers only the hard samples mined in the first stage is utilized for the fusion network. By doing so, complex samples that could not be represented by one type of feature could be focused and the model could be forced to exploit other domain-related information more effectively. Experiment results show that the proposed model outperforms state-of-the-art models on both the National Institute of Health (NIH) and our CURE database.
****
Learn to Augment: Joint Data Augmentation and Network Optimization for Text Recognition
Canjie Luo, Yuanzhi Zhu, Lianwen Jin, Yongpan Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13746-13755
Handwritten text and scene text suffer from various shapes and distorted patterns. Thus training a robust recognition model requires a large amount of data to cover diversity as much as possible. In contrast to data collection and annotation, data augmentation is a low cost way. In this paper, we propose a new method for text image augmentation. Different from traditional augmentation methods such as rotation, scaling and perspective transformation, our proposed augmentation

method is designed to learn proper and efficient data augmentation which is more effective and specific for training a robust recognizer. By using a set of custom fiducial points, the proposed augmentation method is flexible and controllable. Furthermore, we bridge the gap between the isolated processes of data augmentation and network optimization by joint learning. An agent network learns from the output of the recognition network and controls the fiducial points to generate more proper training samples for the recognition network. Extensive experiments on various benchmarks, including regular scene text, irregular scene text and handwritten text, show that the proposed augmentation and the joint learning methods significantly boost the performance of the recognition networks. A general toolkit for geometric augmentation is available.

****

PointGMM: A Neural GMM Network for Point Clouds

Amir Hertz, Rana Hanocka, Raja Giryes, Daniel Cohen-Or; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12054-12063

Point clouds are a popular representation for 3D shapes. However, they encode a particular sampling without accounting for shape priors or non-local information. We advocate for the use of a hierarchical Gaussian mixture model (hGMM), which is a compact, adaptive and lightweight representation that probabilistically defines the underlying 3D surface. We present PointGMM, a neural network that learns to generate hGMMs which are characteristic of the shape class, and also coincide with the input point cloud. PointGMM is trained over a collection of shapes to learn a class-specific prior. The hierarchical representation has two main advantages: (i) coarse-to-fine learning, which avoids converging to poor local-minima; and (ii) (an unsupervised) consistent partitioning of the input shape. We show that as a generative model, PointGMM learns a meaningful latent space which enables generating consistent interpolations between existing shapes, as well as synthesizing novel shapes. We also present a novel framework for rigid registration using PointGMM, that learns to disentangle orientation from structure of an input shape.

****

Weakly Supervised Semantic Point Cloud Segmentation: Towards 10x Fewer Labels

Xun Xu, Gim Hee Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13706-13715

Point cloud analysis has received much attention recently; and segmentation is one of the most important tasks. The success of existing approaches is attributed to deep network design and large amount of labelled training data, where the latter is assumed to be always available. However, obtaining 3d point cloud segmentation labels is often very costly in practice. In this work, we propose a weakly supervised point cloud segmentation approach which requires only a tiny fraction of points to be labelled in the training stage. This is made possible by learning gradient approximation and exploitation of additional spatial and color smoothness constraints. Experiments are done on three public datasets with different degrees of weak supervision. In particular, our proposed method can produce results that are close to and sometimes even better than its fully supervised counterpart with 10X fewer labels.

****

CoverNet: Multimodal Behavior Prediction Using Trajectory Sets

Tung Phan-Minh, Elena Corina Grigore, Freddy A. Boulton, Oscar Beijbom, Eric M. Wolff; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14074-14083

We present CoverNet, a new method for multimodal, probabilistic trajectory prediction for urban driving. Previous work has employed a variety of methods, including multimodal regression, occupancy maps, and 1-step stochastic policies. We instead frame the trajectory prediction problem as classification over a diverse set of trajectories. The size of this set remains manageable due to the limited number of distinct actions that can be taken over a reasonable prediction horizon. We structure the trajectory set to a) ensure a desired level of coverage of the state space, and b) eliminate physically impossible trajectories. By dynamical

ly generating trajectory sets based on the agent's current state, we can further improve our method's efficiency. We demonstrate our approach on public, real world self-driving datasets, and show that it outperforms state-of-the-art methods.

****

Screencast Tutorial Video Understanding

Kunpeng Li, Chen Fang, Zhaowen Wang, Seokhwan Kim, Hailin Jin, Yun Fu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12526-12535

Screencast tutorials are videos created by people to teach how to use software applications or demonstrate procedures for accomplishing tasks. It is very popular for both novice and experienced users to learn new skills, compared to other tutorial media such as text, because of the visual guidance and the ease of understanding. In this paper, we propose visual understanding of screencast tutorials as a new research problem to the computer vision community. We collect a new dataset of Adobe Photoshop video tutorials and annotate it with both low-level and high-level semantic labels. We introduce a bottom-up pipeline to understand Photoshop video tutorials. We leverage state-of-the-art object detection algorithms with domain specific visual cues to detect important events in a video tutorial and segment it into clips according to the detected events. We propose a visual cue reasoning algorithm for two high-level tasks: video retrieval and video captioning. We conduct extensive evaluations of the proposed pipeline. Experimental results show that it is effective in terms of understanding video tutorials. We believe our work will serves as a starting point for future research on this important application domain of video understanding.

****

Gated Channel Transformation for Visual Recognition

Zongxin Yang, Linchao Zhu, Yu Wu, Yi Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11794-11803

In this work, we propose a generally applicable transformation unit for visual recognition with deep convolutional neural networks. This transformation explicitly models channel relationships with explainable control variables. These variables determine the neuron behaviors of competition or cooperation, and they are jointly optimized with the convolutional weight towards more accurate recognition. In Squeeze-and-Excitation (SE) Networks, the channel relationships are implicitly learned by fully connected layers, and the SE block is integrated at the block-level. We instead introduce a channel normalization layer to reduce the number of parameters and computational complexity. This lightweight layer incorporates a simple l2 normalization, enabling our transformation unit applicable to operator-level without much increase of additional parameters. Extensive experiments demonstrate the effectiveness of our unit with clear margins on many vision tasks, i.e., image classification on ImageNet, object detection and instance segmentation on COCO, video classification on Kinetics.

****

Learning to Measure the Static Friction Coefficient in Cloth Contact

Abdullah Haroon Rasheed, Victor Romero, Florence Bertails-Descoubes, Stefanie Wuhrer, Jean-Sebastien Franco, Arnaud Lazarus; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9912-9921

Measuring friction coefficients between cloth and an external body is a longstanding issue in mechanical engineering, never yet addressed with a pure vision-based system. The latter offers the prospect of simpler, less invasive friction measurement protocols compared to traditional ones, and can vastly benefit from recent deep learning advances. Such a novel measurement strategy however proves challenging, as no large labelled dataset for cloth contact exists, and creating one would require thousands of physics workbench measurements with broad coverage of cloth-material pairs. Using synthetic data instead is only possible assuming the availability of a soft-body mechanical simulator with true-to-life friction physics accuracy, yet to be verified. We propose a first vision-based measurement network for friction between cloth and a substrate, using a simple and repeatable video acquisition protocol. We train our network on purely synthetic data ge

nerated by a state-of-the-art frictional contact simulator, which we carefully c
alibrate and validate against real experiments under controlled conditions. We s
how promising results on a large set of contact pairs between real cloth samples
 and various kinds of substrates, with 93.6% of all measurements predicted withi
n 0.1 range of standard physics bench measurements.
****

Can Deep Learning Recognize Subtle Human Activities?
Vincent Jacquot, Zhuofan Ying, Gabriel Kreiman; Proceedings of the IEEE/CVF Co
nference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14244-1425
3
Deep Learning has driven recent and exciting progress in computer vision, instil
ling the belief that these algorithms could solve any visual task. Yet, datasets
 commonly used to train and test computer vision algorithms have pervasive confo
unding factors. Such biases make it difficult to truly estimate the performance
of those algorithms and how well computer vision models can extrapolate outside
the distribution in which they were trained. In this work, we propose a new acti
on classification challenge that is performed well by humans, but poorly by stat
e-of-the-art Deep Learning models. As a proof-of-principle, we consider three ex
emplary tasks: drinking, reading, and sitting. The best accuracies reached using
 state-of-the-art computer vision models were 61.7%, 62.8%, and 76.8%, respectiv
ely, while human participants scored above 90% accuracy on the three tasks. We p
ropose a rigorous method to reduce confounds when creating datasets, and when co
mparing human versus computer vision performance. Source code and datasets are p
ublicly available.
****

Combating Noisy Labels by Agreement: A Joint Training Method with Co-Regularizat
ion
Hongxin Wei, Lei Feng, Xiangyu Chen, Bo An; Proceedings of the IEEE/CVF Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13726-13735
Deep Learning with noisy labels is a practically challenging problem in weakly-s
upervised learning. The state-of-the-art approaches "Decoupling" and "Co-teachin
g+" claim that the "disagreement" strategy is crucial for alleviating the proble
m of learning with noisy labels. In this paper, we start from a different perspe
ctive and propose a robust learning paradigm called JoCoR, which aims to reduce
the diversity of two networks during training. Specifically, we first use two ne
tworks to make predictions on the same mini-batch data and calculate a joint los
s with Co-Regularization for each training example. Then we select small-loss ex
amples to update the parameters of both two networks simultaneously. Trained by
the joint loss, these two networks would be more and more similar due to the eff
ect of Co-Regularization. Extensive experimental results on corrupted data from
benchmark datasets including MNIST, CIFAR-10, CIFAR-100 and Clothing1M demonstra
te that JoCoR is superior to many state-of-the-art approaches for learning with
noisy labels.
****

Superpixel Segmentation With Fully Convolutional Networks
Fengting Yang, Qian Sun, Hailin Jin, Zihan Zhou; Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13964-13
973
In computer vision, superpixels have been widely used as an effective way to red
uce the number of image primitives for subsequent processing. But only a few att
empts have been made to incorporate them into deep neural networks. One main rea
son is that the standard convolution operation is defined on regular grids and b
ecomes inefficient when applied to superpixels. Inspired by an initialization st
rategy commonly adopted by traditional superpixel algorithms, we present a novel
 method that employs a simple fully convolutional network to predict superpixels
 on a regular image grid. Experimental results on benchmark datasets show that o
ur method achieves state-of-the-art superpixel segmentation performance while ru
nning at about 50fps. Based on the predicted superpixels, we further develop a d
ownsampling/upsampling scheme for deep networks with the goal of generating high
-resolution outputs for dense prediction tasks. Specifically, we modify a popula

r network architecture for stereo matching to simultaneously predict superpixels and disparities. We show that improved disparity estimation accuracy can be obtained on public datasets.
****

ContourNet: Taking a Further Step Toward Accurate Arbitrary-Shaped Scene Text Detection
Yuxin Wang, Hongtao Xie, Zheng-Jun Zha, Mengting Xing, Zilong Fu, Yongdong Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11753-11762
Scene text detection has witnessed rapid development in recent years. However, there still exists two main challenges: 1) many methods suffer from false positives in their text representations; 2) the large scale variance of scene texts makes it hard for network to learn samples. In this paper, we propose the ContourNet, which effectively handles these two problems taking a further step toward accurate arbitrary-shaped text detection. At first, a scale-insensitive Adaptive Region Proposal Network (Adaptive-RPN) is proposed to generate text proposals by only focusing on the Intersection over Union (IoU) values between predicted and ground-truth bounding boxes. Then a novel Local Orthogonal Texture-aware Module (LOTM) models the local texture information of proposal features in two orthogonal directions and represents text region with a set of contour points. Considering that the strong unidirectional or weakly orthogonal activation is usually caused by the monotonous texture characteristic of false-positive patterns (e.g. streaks.), our method effectively suppresses these false positives by only outputting predictions with high response value in both orthogonal directions. This gives more accurate description of text regions. Extensive experiments on three challenging datasets (Total-Text, CTW1500 and ICDAR2015) verify that our method achieves the state-of-the-art performance. Code is available at https://github.com/wangyuxin87/ContourNet.
****

Optimal least-squares solution to the hand-eye calibration problem
Amit Dekel, Linus Harenstam-Nielsen, Sergio Caccamo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13598-13606
We propose a least-squares formulation to the noisy hand-eye calibration problem using dual-quaternions, and introduce efficient algorithms to find the exact optimal solution, based on analytic properties of the problem, avoiding non-linear optimization. We further present simple analytic approximate solutions which provide remarkably good estimations compared to the exact solution. In addition, we show how to generalize our solution to account for a given extrinsic prior in the cost function. To the best of our knowledge our algorithm is the most efficient approach to optimally solve the hand-eye calibration problem.
****

Uncertainty-Aware CNNs for Depth Completion: Uncertainty from Beginning to End
Abdelrahman Eldesokey, Michael Felsberg, Karl Holmquist, Michael Persson; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12014-12023
The focus in deep learning research has been mostly to push the limits of prediction accuracy. However, this was often achieved at the cost of increased complexity, raising concerns about the interpretability and the reliability of deep networks. Recently, an increasing attention has been given to untangling the complexity of deep networks and quantifying their uncertainty for different computer vision tasks. Differently, the task of depth completion has not received enough attention despite the inherent noisy nature of depth sensors. In this work, we thus focus on modeling the uncertainty of depth data in depth completion starting from the sparse noisy input all the way to the final prediction. We propose a novel approach to identify disturbed measurements in the input by learning an input confidence estimator in a self-supervised manner based on the normalized convolutional neural networks (NCNNs). Further, we propose a probabilistic version of NCNNs that produces a statistically meaningful uncertainty measure for the final prediction. When we evaluate our approach on the KITTI dataset for depth compl

etion, we outperform all the existing Bayesian Deep Learning approaches in terms of prediction accuracy, quality of the uncertainty measure, and the computational efficiency. Moreover, our small network with 670k parameters performs on-par with conventional approaches with millions of parameters. These results give strong evidence that separating the network into parallel uncertainty and prediction streams leads to state-of-the-art performance with accurate uncertainty estimates.
****

Learning From Web Data With Self-Organizing Memory Module
Yi Tu, Li Niu, Junjie Chen, Dawei Cheng, Liqing Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12846-12855
Learning from web data has attracted lots of research interest in recent years. However, crawled web images usually have two types of noises, label noise and background noise, which induce extra difficulties in utilizing them effectively. Most existing methods either rely on human supervision or ignore the background noise. In this paper, we propose a novel method, which is capable of handling these two types of noises together, without the supervision of clean images in the training stage. Particularly, we formulate our method under the framework of multi-instance learning by grouping ROIs (i.e., images and their region proposals) from the same category into bags. ROIs in each bag are assigned with different weights based on the representative/discriminative scores of their nearest clusters, in which the clusters and their scores are obtained via our designed memory module. Our memory module could be naturally integrated with the classification module, leading to an end-to-end trainable system. Extensive experiments on four benchmark datasets demonstrate the effectiveness of our method.
****

Overcoming Classifier Imbalance for Long-Tail Object Detection With Balanced Group Softmax
Yu Li, Tao Wang, Bingyi Kang, Sheng Tang, Chunfeng Wang, Jintao Li, Jiashi Feng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10991-11000
Solving long-tail large vocabulary object detection with deep learning based models is a challenging and demanding task, which is however under-explored. In this work, we provide the first systematic analysis on the underperformance of state-of-the-art models in front of long-tail distribution. We find existing detection methods are unable to model few-shot classes when the dataset is extremely skewed, which can result in classifier imbalance in terms of parameter magnitude. Directly adapting long-tail classification models to detection frameworks can not solve this problem due to the intrinsic difference between detection and classification. In this work, we propose a novel balanced group softmax (BAGS) module for balancing the classifiers within the detection frameworks through group-wise training. It implicitly modulates the training process for the head and tail classes and ensures they are both sufficiently trained, without requiring any extra sampling for the instances from the tail classes. Extensive experiments on the very recent long-tail large vocabulary object recognition benchmark LVIS show that our proposed BAGS significantly improves the performance of detectors with various backbones and frameworks on both object detection and instance segmentation. It beats all state-of-the-art methods transferred from long-tail image classification and establishes new state-of-the-art. Code is available at https://github.com/FishYuLi/BalancedGroupSoftmax.
****

Hierarchical Scene Coordinate Classification and Regression for Visual Localization
Xiaotian Li, Shuzhe Wang, Yi Zhao, Jakob Verbeek, Juho Kannala; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11983-11992
Visual localization is critical to many applications in computer vision and robotics. To address single-image RGB localization, state-of-the-art feature-based methods match local descriptors between a query image and a pre-built 3D model. R

ecently, deep neural networks have been exploited to regress the mapping between raw pixels and 3D coordinates in the scene, and thus the matching is implicitly performed by the forward pass through the network. However, in a large and ambiguous environment, learning such a regression task directly can be difficult for a single network. In this work, we present a new hierarchical scene coordinate network to predict pixel scene coordinates in a coarse-to-fine manner from a single RGB image. The network consists of a series of output layers, each of them conditioned on the previous ones. The final output layer predicts the 3D coordinates and the others produce progressively finer discrete location labels. The proposed method outperforms the baseline regression-only network and allows us to train compact models which scale robustly to large environments. It sets a new state-of-the-art for single-image RGB localization performance on the 7-Scenes, 12-Scenes, Cambridge Landmarks datasets, and three combined scenes. Moreover, for large-scale outdoor localization on the Aachen Day-Night dataset, we present a hybrid approach which outperforms existing scene coordinate regression methods, and reduces significantly the performance gap w.r.t. explicit feature matching methods.

****

Symmetry and Group in Attribute-Object Compositions

Yong-Lu Li, Yue Xu, Xiaohan Mao, Cewu Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11316-11325

Attributes and objects can compose diverse compositions. To model the compositional nature of these general concepts, it is a good choice to learn them through transformations, such as coupling and decoupling. However, complex transformations need to satisfy specific principles to guarantee the rationality. In this paper, we first propose a previously ignored principle of attribute-object transformation: Symmetry. For example, coupling peeled-apple with attribute peeled should result in peeled-apple, and decoupling peeled from apple should still output apple. Incorporating the symmetry principle, a transformation framework inspired by group theory is built, i.e. SymNet. SymNet consists of two modules, Coupling Network and Decoupling Network. With the group axioms and symmetry property as objectives, we adopt Deep Neural Networks to implement SymNet and train it in an end-to-end paradigm. Moreover, we propose a Relative Moving Distance (RMD) based recognition method to utilize the attribute change instead of the attribute pattern itself to classify attributes. Our symmetry learning can be utilized for the Compositional Zero-Shot Learning task and outperforms the state-of-the-art on widely-used benchmarks. Code is available at https://github.com/DirtyHarryLYL/SymNet.

****

SurfelGAN: Synthesizing Realistic Sensor Data for Autonomous Driving

Zhenpei Yang, Yuning Chai, Dragomir Anguelov, Yin Zhou, Pei Sun, Dumitru Erhan, Sean Rafferty, Henrik Kretzschmar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11118-11127

Autonomous driving system development is critically dependent on the ability to replay complex and diverse traffic scenarios in simulation. In such scenarios, the ability to accurately simulate the vehicle sensors such as cameras, lidar or radar is hugely helpful. However, current sensor simulators leverage gaming engines such as Unreal or Unity, requiring manual creation of environments, objects, and material properties. Such approaches have limited scalability and fail to produce realistic approximations of camera, lidar, and radar data without significant additional work. In this paper, we present a simple yet effective approach to generate realistic scenario sensor data, based only on a limited amount of lidar and camera data collected by an autonomous vehicle. Our approach uses texture-mapped surfels to efficiently reconstruct the scene from an initial vehicle pass or set of passes, preserving rich information about object 3D geometry and appearance, as well as the scene conditions. We then leverage a SurfelGAN network to reconstruct realistic camera images for novel positions and orientations of the self-driving vehicle and moving objects in the scene. We demonstrate our approach on the Waymo Open Dataset and show that it can synthesize realistic camera data for simulated scenarios. We also create a novel dataset that contains cases

in which two self-driving vehicles observe the same scene at the same time. We use this dataset to provide additional evaluation and demonstrate the usefulness of our SurfelGAN model.
****

What Machines See Is Not What They Get: Fooling Scene Text Recognition Models With Adversarial Text Images
Xing Xu, Jiefu Chen, Jinhui Xiao, Lianli Gao, Fumin Shen, Heng Tao Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12304-12314
The research on scene text recognition (STR) has made remarkable progress in recent years with the development of deep neural networks (DNNs). Recent studies on adversarial attack have verified that a DNN model designed for non-sequential tasks (e.g., classification, segmentation and retrieval) can be easily fooled by adversarial examples. Actually, STR is an application highly related to security issues. However, there are few studies considering the safety and reliability of STR models that make sequential prediction. In this paper, we make the first attempt in attacking the state-of-the-art DNN-based STR models. Specifically, we propose a novel and efficient optimization-based method that can be naturally integrated to different sequential prediction schemes, i.e., connectionist temporal classification (CTC) and attention mechanism. We apply our proposed method to five state-of-the-art STR models with both targeted and untargeted attack modes, the comprehensive results on 7 real-world datasets and 2 synthetic datasets consistently show the vulnerability of these STR models with a significant performance drop. Finally, we also test our attack method on a real-world STR engine of Baidu OCR, which demonstrates the practical potentials of our method.
****

Learning to Learn Single Domain Generalization
Fengchun Qiao, Long Zhao, Xi Peng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12556-12565
We are concerned with a worst-case scenario in model generalization, in the sense that a model aims to perform well on many unseen domains while there is only one single domain available for training. We propose a new method named adversarial domain augmentation to solve this Out-of-Distribution (OOD) generalization problem. The key idea is to leverage adversarial training to create "fictitious" yet "challenging" populations, from which a model can learn to generalize with theoretical guarantees. To facilitate fast and desirable domain augmentation, we cast the model training in a meta-learning scheme and use a Wasserstein Auto-Encoder (WAE) to relax the widely used worst-case constraint. Detailed theoretical analysis is provided to testify our formulation, while extensive experiments on multiple benchmark datasets indicate its superior performance in tackling single domain generalization.
****

Warp to the Future: Joint Forecasting of Features and Feature Motion
Josip Saric, Marin Orsic, Tonci Antunovic, Sacha Vrazic, Sinisa Segvic; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10648-10657
We address anticipation of scene development by forecasting semantic segmentation of future frames. Several previous works approach this problem by F2F (feature-to-feature) forecasting where future features are regressed from observed features. Different from previous work, we consider a novel F2M (feature-to-motion) formulation, which performs the forecast by warping observed features according to regressed feature flow. This formulation models a causal relationship between the past and the future, and regularizes inference by reducing dimensionality of the forecasting target. However, emergence of future scenery which was not visible in observed frames can not be explained by warping. We propose to address this issue by complementing F2M forecasting with the classic F2F approach. We realize this idea as a multi-head F2MF model built atop shared features. Experiments show that the F2M head prevails in static parts of the scene while the F2F head kicks-in to fill-in the novel regions. The proposed F2MF model operates in synergy with correlation features and outperforms all previous approaches both in sh

ort-term and mid-term forecast on the Cityscapes dataset.
****
Action Genome: Actions As Compositions of Spatio-Temporal Scene Graphs
Jingwei Ji, Ranjay Krishna, Li Fei-Fei, Juan Carlos Niebles; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10236-10247
Action recognition has typically treated actions and activities as monolithic events that occur in videos. However, there is evidence from Cognitive Science and Neuroscience that people actively encode activities into consistent hierarchical part structures. However, in Computer Vision, few explorations on representations that encode event partonomies have been made. Inspired by evidence that the prototypical unit of an event is an action-object interaction, we introduce Action Genome, a representation that decomposes actions into spatio-temporal scene graphs. Action Genome captures changes between objects and their pairwise relationships while an action occurs. It contains 10K videos with 0.4M objects and 1.7M visual relationships annotated. With Action Genome, we extend an existing action recognition model by incorporating scene graphs as spatio-temporal feature banks to achieve better performance on the Charades dataset. Next, by decomposing and learning the temporal changes in visual relationships that result in an action, we demonstrate the utility of a hierarchical event decomposition by enabling few-shot action recognition, achieving 42.7% mAP using as few as 10 examples. Finally, we benchmark existing scene graph models on the new task of spatio-temporal scene graph prediction.
****
Speech2Action: Cross-Modal Supervision for Action Recognition
Arsha Nagrani, Chen Sun, David Ross, Rahul Sukthankar, Cordelia Schmid, Andrew Zisserman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10317-10326
Is it possible to guess human action from dialogue alone? In this work we investigate the link between spoken words and actions in movies. We note that movie screenplays describe actions, as well as contain the speech of characters and hence can be used to learn this correlation with no additional supervision. We train a BERT-based Speech2Action classifier on over a thousand movie screenplays, to predict action labels from transcribed speech segments. We then apply this model to the speech segments of a large unlabelled movie corpus (188M speech segments from 288K movies). Using the predictions of this model, we obtain weak action labels for over 800K video clips. By training on these video clips, we demonstrate superior action recognition performance on standard action recognition benchmarks, without using a single manually labelled action example.
****
Learning to Cluster Faces via Confidence and Connectivity Estimation
Lei Yang, Dapeng Chen, Xiaohang Zhan, Rui Zhao, Chen Change Loy, Dahua Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13369-13378
Face clustering is an essential tool for exploiting the unlabeled face data, and has a wide range of applications including face annotation and retrieval. Recent works show that supervised clustering can result in noticeable performance gain. However, they usually involve heuristic steps and require numerous overlapped subgraphs, severely restricting their accuracy and efficiency. In this paper, we propose a fully learnable clustering framework without requiring a large number of overlapped subgraphs. Instead, we transform the clustering problem into two sub-problems. Specifically, two graph convolutional networks, named GCN-V and GCN-E, are designed to estimate the confidence of vertices and the connectivity of edges, respectively. With the vertex confidence and edge connectivity, we can naturally organize more relevant vertices on the affinity graph and group them into clusters. Experiments on two large-scale benchmarks show that our method significantly improves clustering accuracy and thus performance of the recognition models trained on top, yet it is an order of magnitude more efficient than existing supervised methods.
****

Rethinking Performance Estimation in Neural Architecture Search
Xiawu Zheng, Rongrong Ji, Qiang Wang, Qixiang Ye, Zhenguo Li, Yonghong Tian , Qi Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Patter n Recognition (CVPR), 2020, pp. 11356-11365
Neural architecture search (NAS) remains a challenging problem, which is attribu ted to the indispensable and time-consuming component of performance estimation (PE). In this paper, we provide a novel yet systematic rethinking of PE in a res ource constrained regime, termed budgeted PE (BPE), which precisely and effectiv ely estimates the performance of an architecture sampled from an architecture sp ace. Since searching an optimal BPE is extremely time-consuming as it requires t o train a large number of networks for evaluation, we propose a Minimum Importan ce Pruning (MIP) approach. Given a dataset and a BPE search space, MIP estimates the importance of hyper-parameters using random forest and subsequently prunes the minimum one from the next iteration. In this way, MIP effectively prunes les s important hyper-parameters to allocate more computational resource on more imp ortant ones, thus achieving an effective exploration. By combining BPE with vari ous search algorithms including reinforcement learning, evolution algorithm, ran dom search, and differentiable architecture search, we achieve 1, 000x of NAS sp eed up with a negligible performance drop comparing to the SOTA. All the NAS sea rch codes are available at: https: //github.com/zhengxiawu/rethinking_performanc e_ estimation_in_NAS
****
Revisiting the Sibling Head in Object Detector
Guanglu Song, Yu Liu, Xiaogang Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11563-11572
The "shared head for classification and localization" (sibling head), firstly de nominated in Fast RCNN, has been leading the fashion of the object detection com munity in the past five years. This paper provides the observation that the spat ial misalignment between the two object functions in the sibling head can consid erably hurt the training process, but this misalignment can be resolved by a ver y simple operator called task-aware spatial disentanglement (TSD). Considering t he classification and regression, TSD decouples them from the spatial dimension by generating two disentangled proposals for them, which are estimated by the sh ared proposal. This is inspired by the natural insight that for one instance, th e features in some salient area may have rich information for classification whi le these around the boundary may be good at bounding box regression. Surprisingl y, this simple design can boost all backbones and models on both MS COCO and Goo gle OpenImage consistently by 3% mAP. Further, we propose a progressive constra int to enlarge the performance margin between the disentangled and the shared pr oposals, and gain 1% more mAP. We show the TSD breaks through the upper bound o f nowadays single-model detector by a large margin (mAP 49.4 with ResNet-101, 51 .2 with SENet154), and is the core model of our 1st place solution on the Google OpenImage Challenge 2019.
****
EcoNAS: Finding Proxies for Economical Neural Architecture Search
Dongzhan Zhou, Xinchi Zhou, Wenwei Zhang, Chen Change Loy, Shuai Yi, Xuesen Zhang, Wanli Ouyang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11396-11404
Neural Architecture Search (NAS) achieves significant progress in many computer vision tasks. While many methods are proposed to improve the efficiency of NAS, the search progress is still laborious because training and evaluating plausible architectures over large search space is time-consuming. Assessing network cand idates under a proxy (i.e., computationally reduced setting) thus becomes inevit able. In this paper, we observe that most existing proxies exhibit different beh aviors in maintaining the rank consistency among network candidates. In particul ar, some proxies can be more reliable - the rank of candidates does not differ m uch comparing their reduced setting performance and final performance. In this p aper, we systematically investigate some widely adopted reduction factors and re port our observations. Inspired by these observations, we present a reliable pro xy and further formulate a hierarchical proxy strategy that spends more computat

ions on candidate networks that are potentially more accurate, while discards unpromising ones in early stage with a fast proxy. This leads to an economical evolutionary-based NAS (EcoNAS), which achieves an impressive 400xsearch time reduction in comparison to the evolutionary-based state of the art [19] (8 v.s. 3150 GPU days). Some new proxies led by our observations can also be applied to accelerate other NAS methods while still able to discover good candidate networks with performance matching those found by previous proxy strategies. Codes and models will be released to facilitate future research.
****

Norm-Aware Embedding for Efficient Person Search
Di Chen, Shanshan Zhang, Jian Yang, Bernt Schiele; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12615-12624
Person Search is a practically relevant task that aims to jointly solve Person Detection and Person Re-identification (re-ID). Specifically, it requires to find and locate all instances with the same identity as the query person in a set of panoramic gallery images. One major challenge comes from the contradictory goals of the two sub-tasks, i.e., person detection focuses on finding the commonness of all persons while person re-ID handles the differences among multiple identities. Therefore, it is crucial to reconcile the relationship between the two sub-tasks in a joint person search model. To this end, We present a novel approach called Norm-Aware Embedding to disentangle the person embedding into norm and angle for detection and re-ID respectively, allowing for both effective and efficient multi-task training. We further extend the proposal-level person embedding to pixel-level, whose discrimination ability is less affected by mis-alignment. We outperform other one-step methods by a large margin and achieve comparable performance to two-step methods on both CUHK-SYSU and PRW. Also, Our method is easy to train and resource-friendly, running at 12 fps on a single GPU.
****

Syntax-Aware Action Targeting for Video Captioning
Qi Zheng, Chaoyue Wang, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13096-13105
Existing methods on video captioning have made great efforts to identify objects/instances in videos, but few of them emphasize the prediction of action. As a result, the learned models are likely to depend heavily on the prior of training data, such as the co-occurrence of objects, which may cause an enormous divergence between the generated descriptions and the video content. In this paper, we explicitly emphasize the importance of action by predicting visually-related syntax components including subject, object and predicate. Specifically, we propose a Syntax-Aware Action Targeting (SAAT) module that firstly builds a self-attended scene representation to draw global dependence among multiple objects within a scene, and then decodes the visually-related syntax components by setting different queries. After targeting the action, indicated by predicate, our captioner learns an attention distribution over the predicate and the previously predicted words to guide the generation of the next word. Comprehensive experiments on MSVD and MSR-VTT datasets demonstrate the efficacy of the proposed model.
****

On Vocabulary Reliance in Scene Text Recognition
Zhaoyi Wan, Jielei Zhang, Liang Zhang, Jiebo Luo, Cong Yao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11425-11434
The pursuit of high performance on public benchmarks has been the driving force for research in scene text recognition, and notable progresses have been achieved. However, a close investigation reveals a startling fact that the state-of-the-art methods perform well on images with words within vocabulary but generalize poorly to images with words outside vocabulary. We call this phenomenon "vocabulary reliance". In this paper, we establish an analytical framework, in which different datasets, metrics and module combinations for quantitative comparisons are devised, to conduct an in-depth study on the problem of vocabulary reliance in scene text recognition. Key findings include: (1) Vocabulary reliance is ubiqui

tous, i.e., all existing algorithms more or less exhibit such characteristic; (2) Attention-based decoders prove weak in generalizing to words outside vocabulary and segmentation-based decoders perform well in utilizing visual features; (3) Context modeling is highly coupled with the prediction layers. These findings provide new insights and can benefit future research in scene text recognition. Furthermore, we propose a simple yet effective mutual learning strategy to allow models of two families (attention-based and segmentation-based) to learn collaboratively. This remedy alleviates the problem of vocabulary reliance and significantly improves the overall scene text recognition performance.
****

Imitative Non-Autoregressive Modeling for Trajectory Forecasting and Imputation
Mengshi Qi, Jie Qin, Yu Wu, Yi Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12736-12745
Trajectory forecasting and imputation are pivotal steps towards understanding the movement of human and objects, which are quite challenging since the future trajectories and missing values in a temporal sequence are full of uncertainties, and the spatial-temporally contextual correlation is hard to model. Yet, the relevance between sequence prediction and imputation is disregarded by existing approaches. To this end, we propose a novel imitative non-autoregressive modeling method to simultaneously handle the trajectory prediction task and the missing value imputation task. Specifically, our framework adopts an imitation learning paradigm, which contains a recurrent conditional variational autoencoder (RC-VAE) as a demonstrator, and a non-autoregressive transformation model (NART) as a learner. By jointly optimizing the two models, RC-VAE can predict the future trajectory and capture the temporal relationship in the sequence to supervise the NART learner. As a result, NART learns from the demonstrator and imputes the missing value in a non autoregressive strategy. We conduct extensive experiments on three popular datasets, and the results show that our model achieves state-of-the-art performance across all the datasets.
****

Hi-CMD: Hierarchical Cross-Modality Disentanglement for Visible-Infrared Person Re-Identification
Seokeon Choi, Sumin Lee, Youngeun Kim, Taekyung Kim, Changick Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10257-10266
Visible-infrared person re-identification (VI-ReID) is an important task in night-time surveillance applications, since visible cameras are difficult to capture valid appearance information under poor illumination conditions. Compared to traditional person re-identification that handles only the intra-modality discrepancy, VI-ReID suffers from additional cross-modality discrepancy caused by different types of imaging systems. To reduce both intra- and cross-modality discrepancies, we propose a Hierarchical Cross-Modality Disentanglement (Hi-CMD) method, which automatically disentangles ID-discriminative factors and ID-excluded factors from visible-thermal images. We only use ID-discriminative factors for robust cross-modality matching without ID-excluded factors such as pose or illumination. To implement our approach, we introduce an ID-preserving person image generation network and a hierarchical feature learning module. Our generation network learns the disentangled representation by generating a new cross-modality image with different poses and illuminations while preserving a person's identity. At the same time, the feature learning module enables our model to explicitly extract the common ID-discriminative characteristic between visible-infrared images. Extensive experimental results demonstrate that our method outperforms the state-of-the-art methods on two VI-ReID datasets. The source code is available at: https://github.com/bismex/HiCMD.
****

Say As You Wish: Fine-Grained Control of Image Caption Generation With Abstract Scene Graphs
Shizhe Chen, Qin Jin, Peng Wang, Qi Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9962-9971
Humans are able to describe image contents with coarse to fine details as they w

ish. However, most image captioning models are intention-agnostic which cannot g
enerate diverse descriptions according to different user intentions initiatively
. In this work, we propose the Abstract Scene Graph (ASG) structure to represent
 user intention in fine-grained level and control what and how detailed the gene
rated description should be. The ASG is a directed graph consisting of three typ
es of abstract nodes (object, attribute, relationship) grounded in the image wit
hout any concrete semantic labels. Thus it is easy to obtain either manually or
automatically. From the ASG, we propose a novel ASG2Caption model, which is able
 to recognise user intentions and semantics in the graph, and therefore generate
 desired captions following the graph structure. Our model achieves better contr
ollability conditioning on ASGs than carefully designed baselines on both Visual
Genome and MSCOCO datasets. It also significantly improves the caption diversity
 via automatically sampling diverse ASGs as control signals. Code will be releas
ed at https://github.com/cshizhe/asg2cap.
****

TESA: Tensor Element Self-Attention via Matricization
Francesca Babiloni, Ioannis Marras, Gregory Slabaugh, Stefanos Zafeiriou; Pro
ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (
CVPR), 2020, pp. 13945-13954
Representation learning is a fundamental part of modern computer vision, where a
bstract representations of data are encoded as tensors optimized to solve proble
ms like image segmentation and inpainting. Recently, self-attention in the form
of Non-Local Block has emerged as a powerful technique to enrich features, by ca
pturing complex interdependencies in feature tensors. However, standard self-att
ention approaches leverage only spatial relationships, drawing similarities betw
een vectors and overlooking correlations between channels. In this paper, we int
roduce a new method, called Tensor Element Self-Attention (TESA) that generalize
s such work to capture interdependencies along all dimensions of the tensor usin
g matricization. An order R tensor produces R results, one for each dimension. T
he results are then fused to produce an enriched output which encapsulates simil
arity among tensor elements. Additionally, we analyze self-attention mathematica
lly, providing new perspectives on how it adjusts the singular values of the inp
ut feature tensor. With these new insights, we present experimental results demo
nstrating how TESA can benefit diverse problems including classification and ins
tance segmentation. By simply adding a TESA module to existing networks, we subs
tantially improve competitive baselines and set new state-of-the-art results for
 image inpainting on Celeb and low light raw-to-rgb image translation on SID.
****

Clean-Label Backdoor Attacks on Video Recognition Models
Shihao Zhao, Xingjun Ma, Xiang Zheng, James Bailey, Jingjing Chen, Yu-Gang
Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec
ognition (CVPR), 2020, pp. 14443-14452
Deep neural networks (DNNs) are vulnerable to backdoor attacks which can hide ba
ckdoor triggers in DNNs by poisoning training data. A backdoored model behaves n
ormally on clean test images, yet consistently predicts a particular target clas
s for any test examples that contain the trigger pattern. As such, backdoor atta
cks are hard to detect, and have raised severe security concerns in real-world a
pplications. Thus far, backdoor research has mostly been conducted in the image
domain with image classification models. In this paper, we show that existing im
age backdoor attacks are far less effective on videos, and outline 4 strict cond
itions where existing attacks are likely to fail: 1) scenarios with more input d
imensions (eg. videos), 2) scenarios with high resolution, 3) scenarios with a l
arge number of classes and few examples per class (a "sparse dataset"), and 4) a
ttacks with access to correct labels (eg. clean-label attacks). We propose the u
se of a universal adversarial trigger as the backdoor trigger to attack video re
cognition models, a situation where backdoor attacks are likely to be challenged
 by the above 4 strict conditions. We show on benchmark video datasets that our
proposed backdoor attack can manipulate state-of-the-art video models with high
success rates by poisoning only a small proportion of training data (without cha
nging the labels). We also show that our proposed backdoor attack is resistant t

o state-of-the-art backdoor defense/detection methods, and can even be applied t
o improve image backdoor attacks. Our proposed video backdoor attack not only se
rves as a strong baseline for improving the robustness of video models, but also
 provides a new perspective for more understanding more powerful backdoor attack
s.
****

RPM-Net: Robust Point Matching Using Learned Features
Zi Jian Yew, Gim Hee Lee; Proceedings of the IEEE/CVF Conference on Computer Vi
sion and Pattern Recognition (CVPR), 2020, pp. 11824-11833
Iterative Closest Point (ICP) solves the rigid point cloud registration problem
iteratively in two steps: (1) make hard assignments of spatially closest point c
orrespondences, and then (2) find the least-squares rigid transformation. The ha
rd assignments of closest point correspondences based on spatial distances are s
ensitive to the initial rigid transformation and noisy/outlier points, which oft
en cause ICP to converge to wrong local minima. In this paper, we propose the RP
M-Net -- a less sensitive to initialization and more robust deep learning-based
approach for rigid point cloud registration. To this end, our network uses the d
ifferentiable Sinkhorn layer and annealing to get soft assignments of point corr
espondences from hybrid features learned from both spatial coordinates and local
 geometry. To further improve registration performance, we introduce a secondary
 network to predict optimal annealing parameters. Unlike some existing methods,
our RPM-Net handles missing correspondences and point clouds with partial visibi
lity. Experimental results show that our RPM-Net achieves state-of-the-art perfo
rmance compared to existing non-deep learning and recent deep learning methods.
Our source code is available at the project website (https://github.com/yewzijia
n/RPMNet).
****

Improving One-Shot NAS by Suppressing the Posterior Fading
Xiang Li, Chen Lin, Chuming Li, Ming Sun, Wei Wu, Junjie Yan, Wanli Ouyang
; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit
ion (CVPR), 2020, pp. 13836-13845
Neural architecture search (NAS) has demonstrated much success in automatically
designing effective neural network architectures. To improve the efficiency of N
AS, previous approaches adopt weight sharing method to force all models share th
e same set of weights. However, it has been observed that a model performing bet
ter with shared weights does not necessarily perform better when trained alone.
In this paper, we analyse existing weight sharing one-shot NAS approaches from a
 Bayesian point of view and identify the Posterior Fading problem, which comprom
ises the effectiveness of shared weights. To alleviate this problem, we present
a novel approach to guide the parameter posterior towards its true distribution.
 Moreover, a hard latency constraint is introduced during the search so that the
 desired latency can be achieved. The resulted method, namely Posterior Converge
nt NAS (PC-NAS), achieves state-of-the-art performance under standard GPU latenc
y constraint on ImageNet.
****

Understanding Human Hands in Contact at Internet Scale
Dandan Shan, Jiaqi Geng, Michelle Shu, David F. Fouhey; Proceedings of the IE
EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9
869-9878
Hands are the central means by which humans manipulate their world and being abl
e to reliably extract hand state information from Internet videos of humans enga
ged in their hands has the potential to pave the way to systems that can learn f
rom petabytes of video data. This paper proposes steps towards this by inferring
 a rich representation of hands engaged in interaction method that includes: han
d location, side, contact state, and a box around the object in contact. To supp
ort this effort, we gather a large-scale dataset of hands in contact with object
s consisting of 131 days of footage as well as a 100K annotated hand-contact vid
eo frame dataset. The learned model on this dataset can serve as a foundation fo
r hand-contact understanding in videos. We quantitatively evaluate it both on it
s own and in service of predicting and learning from 3D meshes of human hands.

****
Self-Supervised Equivariant Attention Mechanism for Weakly Supervised Semantic S egmentation
Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, Xilin Chen; Proceedings of t he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12275-12284
Image-level weakly supervised semantic segmentation is a challenging problem tha t has been deeply studied in recent years. Most of advanced solutions exploit cl ass activation map (CAM). However, CAMs can hardly serve as the object mask due to the gap between full and weak supervisions. In this paper, we propose a self-supervised equivariant attention mechanism (SEAM) to discover additional supervi sion and narrow the gap. Our method is based on the observation that equivarianc e is an implicit constraint in fully supervised semantic segmentation, whose pix el-level labels take the same spatial transformation as the input images during data augmentation. However, this constraint is lost on the CAMs trained by image -level supervision. Therefore, we propose consistency regularization on predicte d CAMs from various transformed images to provide self-supervision for network l earning. Moreover, we propose a pixel correlation module (PCM), which exploits c ontext appearance information and refines the prediction of current pixel by its similar neighbors, leading to further improvement on CAMs consistency. Extensiv e experiments on PASCAL VOC 2012 dataset demonstrate our method outperforms stat e-of-the-art methods using the same level of supervision. The code is released o nline.
****
TBT: Targeted Neural Network Attack With Bit Trojan
Adnan Siraj Rakin, Zhezhi He, Deliang Fan; Proceedings of the IEEE/CVF Confere nce on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13198-13207
Security of modern Deep Neural Networks (DNNs) is under severe scrutiny as the d eployment of these models become widespread in many intelligence-based applicati ons. Most recently, DNNs are attacked through Trojan which can effectively infec t the model during the training phase and get activated only through specific in put patterns (i.e, trigger) during inference. In this work, for the first time, we propose a novel Targeted Bit Trojan(TBT) method, which can insert a targeted neural Trojan into a DNN through bit-flip attack. Our algorithm efficiently gene rates a trigger specifically designed to locate certain vulnerable bits of DNN w eights stored in main memory (i.e., DRAM). The objective is that once the attack er flips these vulnerable bits, the network still operates with normal inference accuracy with benign input. However, when the attacker activates the trigger by embedding it with any input, the network is forced to classify all inputs to a certain target class. We demonstrate that flipping only several vulnerable bits identified by our method, using available bit-flip techniques (i.e, row-hammer), can transform a fully functional DNN model into a Trojan-infected model. We per form extensive experiments of CIFAR-10, SVHN and ImageNet datasets on both VGG-1 6 and Resnet-18 architectures. Our proposed TBT could classify 92 of test images to a target class with as little as 84 bit-flips out of 88 million weight bits on Resnet-18 for CIFAR10 dataset.
****
End-to-End Learning of Visual Representations From Uncurated Instructional Video s
Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic , Andrew Zisserman; Proceedings of the IEEE/CVF Conference on Computer Vision a nd Pattern Recognition (CVPR), 2020, pp. 9879-9889
Annotating videos is cumbersome, expensive and not scalable. Yet, many strong vi deo models still rely on manually annotated data. With the recent introduction o f the HowTo100M dataset, narrated videos now offer the possibility of learning v ideo representations without manual supervision. In this work we propose a new l earning approach, MIL-NCE, capable of addressing mis- alignments inherent in nar rated videos. With this approach we are able to learn strong video representatio ns from scratch, without the need for any manual annotation. We evaluate our rep resentations on a wide range of four downstream tasks over eight datasets: actio

n recognition (HMDB-51, UCF-101, Kinetics-700), text-to- video retrieval (YouCoo k2, MSR-VTT), action localization (YouTube-8M Segments, CrossTask) and action se gmentation (COIN). Our method outperforms all published self-supervised approach es for these tasks as well as several fully supervised baselines.
****

OrigamiNet: Weakly-Supervised, Segmentation-Free, One-Step, Full Page Text Recog nition by learning to unfold
Mohamed Yousef, Tom E. Bishop; Proceedings of the IEEE/CVF Conference on Comput er Vision and Pattern Recognition (CVPR), 2020, pp. 14710-14719
Text recognition is a major computer vision task with a big set of associated ch allenges. One of those traditional challenges is the coupled nature of text reco gnition and segmentation. This problem has been progressively solved over the pa st decades, going from segmentation based recognition to segmentation free appro aches, which proved more accurate and much cheaper to annotate data for. We take a step from segmentation-free single line recognition towards segmentation-free multi-line / full page recognition. We propose a novel and simple neural networ k module, termed OrigamiNet, that can augment any CTC-trained, fully convolution al single line text recognizer, to convert it into a multi-line version by provi ding the model with enough spatial capacity to be able to properly collapse a 2D input signal into 1D without losing information. Such modified networks can be trained using exactly their same simple original procedure, and using only unseg mented image and text pairs. We carry out a set of interpretability experiments that show that our trained models learn an accurate implicit line segmentation. We achieve state-of-the-art character error rate on both IAM & ICDAR 2017 HTR be nchmarks for handwriting recognition, surpassing all other methods in the litera ture. On IAM we even surpass single line methods that use accurate localization information during training. Our code is available online at https://github.com/ IntuitionMachines/OrigamiNet .
****

Hierarchical Graph Attention Network for Visual Relationship Detection
Li Mi, Zhenzhong Chen; Proceedings of the IEEE/CVF Conference on Computer Visio n and Pattern Recognition (CVPR), 2020, pp. 13886-13895
Visual Relationship Detection (VRD) aims to describe the relationship between tw o objects by providing a structural triplet shown as . Existing graph-based meth ods mainly represent the relationships by an object-level graph, which ignores t o model the triplet-level dependencies. In this work, a Hierarchical Graph Atten tion Network (HGAT) is proposed to capture the dependencies on both object-level and triplet-level. Object-level graph aims to capture the interactions between objects, while the triplet-level graph models the dependencies among relation tr iplets. In addition, prior knowledge and attention mechanism are introduced to f ix the redundant or missing edges on graphs that are constructed according to sp atial correlation. With these approaches, nodes are allowed to attend over their spatial and semantic neighborhoods' features based on the visual or semantic fe ature correlation. Experimental results on the well-known VG and VRD datasets de monstrate that our model significantly outperforms the state-of-the-art methods.
****

Neural Implicit Embedding for Point Cloud Analysis
Kent Fujiwara, Taiichi Hashimoto; Proceedings of the IEEE/CVF Conference on Com puter Vision and Pattern Recognition (CVPR), 2020, pp. 11734-11743
We present a novel representation for point clouds that encapsulates the local c haracteristics of the underlying structure. The key idea is to embed an implicit representation of the point cloud, namely the distance field, into neural netwo rks. One neural network is used to embed a portion of the distance field around a point. The resulting network weights are concatenated to be used as a represen tation of the corresponding point cloud instance. To enable comparison among the weights, Extreme Learning Machine (ELM) is employed as the embedding network. I nvariance to scale and coordinate change can be achieved by introducing a scale commutative activation layer to the ELM, and aligning the distance field into a canonical pose. Experimental results using our representation demonstrate that o ur proposal is capable of similar or better classification and segmentation perf

ormance compared to the state-of-the-art point-based methods, while requiring le ss time for training.
****

Better Captioning With Sequence-Level Exploration
Jia Chen, Qin Jin; Proceedings of the IEEE/CVF Conference on Computer Vision an d Pattern Recognition (CVPR), 2020, pp. 10890-10899
Sequence-level learning objective has been widely used in captioning tasks to ac hieve the state-of-the-art performance for many models. In this objective, the m odel is trained by the reward on the quality of its generated captions (sequence -level). In this work, we show the limitation of the current sequence-level lear ning objective for captioning tasks from both theory and empirical result. In th eory, we show that the current objective is equivalent to only optimizing the pr ecision side of the caption set generated by the model and therefore overlooks t he recall side. Empirical result shows that the model trained by this objective tends to get lower score on the recall side. We propose to add a sequence-level exploration term to the current objective to boost recall. It guides the model t o explore more plausible captions in the training. In this way, the proposed obj ective takes both the precision and recall sides of generated captions into acco unt. Experiments show the effectiveness of the proposed method on both video and image captioning datasets.
****

Moving in the Right Direction: A Regularization for Deep Metric Learning
Deen Dayal Mohan, Nishant Sankaran, Dennis Fedorishin, Srirangaraj Setlur, V enu Govindaraju; Proceedings of the IEEE/CVF Conference on Computer Vision and P attern Recognition (CVPR), 2020, pp. 14591-14599
Deep metric learning leverages carefully designed sampling strategies and loss f unctions that aid in optimizing the generation of a discriminable embedding spac e. While effective sampling of pairs is critical for shaping the metric space du ring training, the relative interactions between pairs, and consequently the for ces exerted on these pairs that direct their displacement in the embedding space can significantly impact the formation of well separated clusters. In this work , we identify a shortcoming of existing loss formulations which fail to consider more optimal directions of pair displacements as another criterion for optimiza tion. We propose a novel direction regularization to explicitly account for the layout of sampled pairs and attempt to introduce orthogonality in the representa tions. The proposed regularization is easily integrated into existing loss funct ions providing considerable performance improvements. We experimentally validate our hypothesis on the Cars-196, CUB-200 and InShop datasets and outperform exis ting methods to yield state-of-the-art results on these datasets.
****

Improved Few-Shot Visual Classification
Peyman Bateni, Raghav Goyal, Vaden Masrani, Frank Wood, Leonid Sigal; Procee dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP R), 2020, pp. 14493-14502
Few-shot learning is a fundamental task in computer vision that carries the prom ise of alleviating the need for exhaustively labeled data. Most few-shot learnin g approaches to date have focused on progressively more complex neural feature e xtractors and classifier adaptation strategies, and the refinement of the task d efinition itself. In this paper, we explore the hypothesis that a simple class-c ovariance-based distance metric, namely the Mahalanobis distance, adopted into a state of the art few-shot learning approach (CNAPS) can, in and of itself, lead to a significant performance improvement. We also discover that it is possible to learn adaptive feature extractors that allow useful estimation of the high di mensional feature covariances required by this metric from surprisingly few samp les. The result of our work is a new "Simple CNAPS" architecture which has up to 9.2% fewer trainable parameters than CNAPS and performs up to 6.1% better than state of the art on the standard few-shot image classification benchmark dataset .
****

Visual Chirality

Zhiqiu Lin, Jin Sun, Abe Davis, Noah Snavely; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12295-12303
How can we tell whether an image has been mirrored? While we understand the geometry of mirror reflections very well, less has been said about how it affects distributions of imagery at scale, despite widespread use for data augmentation in computer vision. In this paper, we investigate how the statistics of visual data are changed by reflection. We refer to these changes as "visual chirality," after the concept of geometric chirality---the notion of objects that are distinct from their mirror image. Our analysis of visual chirality reveals surprising results, including low-level chiral signals pervading imagery stemming from image processing in cameras, to the ability to discover visual chirality in images of people and faces. Our work has implications for data augmentation, self-supervised learning, and image forensics.
****

Neural Architecture Search for Lightweight Non-Local Networks
Yingwei Li, Xiaojie Jin, Jieru Mei, Xiaochen Lian, Linjie Yang, Cihang Xie, Qihang Yu, Yuyin Zhou, Song Bai, Alan L. Yuille; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10297-10306
Non-Local (NL) blocks have been widely studied in various vision tasks. However, it has been rarely explored to embed the NL blocks in mobile neural networks, mainly due to the following challenges: 1) NL blocks generally have heavy computation cost which makes it difficult to be applied in applications where computational resources are limited, and 2) it is an open problem to discover an optimal configuration to embed NL blocks into mobile neural networks. We propose AutoNL to overcome the above two obstacles. Firstly, we propose a Lightweight Non-Local (LightNL) block by squeezing the transformation operations and incorporating compact features. With the novel design choices, the proposed LightNL block is 400 times computationally cheaper than its conventional counterpart without sacrificing the performance. Secondly, by relaxing the structure of the LightNL block to be differentiable during training, we propose an efficient neural architecture search algorithm to learn an optimal configuration of LightNL blocks in an end-to-end manner. Notably, using only 32 GPU hours, the searched AutoNL model achieves 77.7% top-1 accuracy on ImageNet under a typical mobile setting (350M FLOPs), significantly outperforming previous mobile models including MobileNetV2 (+5.7%), FBNet (+2.8%) and MnasNet (+2.1%). Code and models are available at https://github.com/LiYingwei/AutoNL.
****

Private-kNN: Practical Differential Privacy for Computer Vision
Yuqing Zhu, Xiang Yu, Manmohan Chandraker, Yu-Xiang Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11854-11862
With increasing ethical and legal concerns on privacy for deep models in visual recognition, differential privacy has emerged as a mechanism to disguise membership of sensitive data in training datasets. Recent methods like Private Aggregation of Teacher Ensembles (PATE) leverage a large ensemble of teacher models trained on disjoint subsets of private data, to transfer knowledge to a student model with privacy guarantees. However, labeled vision data is often expensive and datasets, when split into many disjoint training sets, lead to significantly sub-optimal accuracy and thus hardly sustain good privacy bounds. We propose a practically data-efficient scheme based on private release of k-nearest neighbor (kNN) queries, which altogether avoids splitting the training dataset. Our approach allows the use of privacy-amplification by subsampling and iterative refinement of the kNN feature embedding. We rigorously analyze the theoretical properties of our method and demonstrate strong experimental performance on practical computer vision datasets for face attribute recognition and person reidentification. In particular, we achieve comparable or better accuracy than PATE while reducing more than 90% of the privacy loss, thereby providing the "most practical method to-date" for private deep learning in computer vision.
****

Old Is Gold: Redefining the Adversarially Learned One-Class Classifier Training Paradigm

Muhammad Zaigham Zaheer, Jin-Ha Lee, Marcella Astrid, Seung-Ik Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14183-14193

A popular method for anomaly detection is to use the generator of an adversarial network to formulate anomaly score over reconstruction loss of input. Due to the rare occurrence of anomalies, optimizing such networks can be a cumbersome task. Another possible approach is to use both generator and discriminator for anomaly detection. However, attributed to the involvement of adversarial training, this model is often unstable in a way that the performance fluctuates drastically with each training step. In this study, we propose a framework that effectively generates stable results across a wide range of training steps and allows us to use both the generator and the discriminator of an adversarial model for efficient and robust anomaly detection. Our approach transforms the fundamental role of a discriminator from identifying real and fake data to distinguishing between good and bad quality reconstructions. To this end, we prepare training examples for the good quality reconstruction by employing the current generator, whereas poor quality examples are obtained by utilizing an old state of the same generator. This way, the discriminator learns to detect subtle distortions that often appear in reconstructions of the anomaly inputs. Extensive experiments performed on Caltech-256 and MNIST image datasets for novelty detection show superior results. Furthermore, on UCSD Ped2 video dataset for anomaly detection, our model achieves a frame-level AUC of 98.1%, surpassing recent state-of-the-art methods

****

Cops-Ref: A New Dataset and Task on Compositional Referring Expression Comprehension

Zhenfang Chen, Peng Wang, Lin Ma, Kwan-Yee K. Wong, Qi Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10086-10095

Referring expression comprehension (REF) aims at identifying a particular object in a scene by a natural language expression. It requires joint reasoning over the textual and visual domains to solve the problem. Some popular referring expression datasets, however, fail to provide an ideal test bed for evaluating the reasoning ability of the models, mainly because 1) their expressions typically describe only some simple distinctive properties of the object and 2) their images contain limited distracting information. To bridge the gap, we propose a new dataset for visual reasoning in context of referring expression comprehension with two main features. First, we design a novel expression engine rendering various reasoning logics that can be flexibly combined with rich visual properties to generate expressions with varying compositionality. Second, to better exploit the full reasoning chain embodied in an expression, we propose a new test setting by adding additional distracting images containing objects sharing similar properties with the referent, thus minimising the success rate of reasoning-free cross-domain alignment. We evaluate several state-of-the-art REF models, but find none of them can achieve promising performance. A proposed modular hard mining strategy performs the best but still leaves substantial room for improvement.

****

Learning Longterm Representations for Person Re-Identification Using Radio Signals

Lijie Fan, Tianhong Li, Rongyao Fang, Rumen Hristov, Yuan Yuan, Dina Katabi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10699-10709

Person Re-Identification (ReID) aims to recognize a person-of-interest across different places and times. Existing ReID methods rely on images or videos collected using RGB cameras. They extract appearance features like clothes, shoes, hair, etc. Such features, however, can change drastically from one day to the next, leading to inability to identify people over extended time periods. In this paper, we introduce RF-ReID, a novel approach that harnesses radio frequency (RF) signals for longterm person ReID. RF signals traverse clothes and reflect off the

human body; thus they can be used to extract more persistent human-identifying features like body size and shape. We evaluate the performance of RF-ReID on longitudinal datasets that span days and weeks, where the person may wear different clothes across days. Our experiments demonstrate that RF-ReID outperforms state-of-the-art RGB-based ReID approaches for long term person ReID. Our results also reveal two interesting features: First since RF signals work in the presence of occlusions and poor lighting, RF-ReID allows for person ReID in such scenarios. Second, unlike photos and videos which reveal personal and private information, RF signals are more privacy-preserving, and hence can help extend person ReID to privacy-concerned domains, like healthcare.
****

DSNAS: Direct Neural Architecture Search Without Parameter Retraining
Shoukang Hu, Sirui Xie, Hehui Zheng, Chunxiao Liu, Jianping Shi, Xunying Liu, Dahua Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12084-12092
If NAS methods are solutions, what is the problem? Most existing NAS methods require two-stage parameter optimization. However, performance of the same architecture in the two stages correlates poorly. In this work, we propose a new problem definition for NAS, task-specific end-to-end, based on this observation. We argue that given a computer vision task for which a NAS method is expected, this definition can reduce the vaguely-defined NAS evaluation to i) accuracy of this task and ii) the total computation consumed to finally obtain a model with satisfying accuracy. Seeing that most existing methods do not solve this problem directly, we propose DSNAS, an efficient differentiable NAS framework that simultaneously optimizes architecture and parameters with a low-biased Monte Carlo estimate. Child networks derived from DSNAS can be deployed directly without parameter retraining. Comparing with two-stage methods, DSNAS successfully discovers networks with comparable accuracy (74.4%) on ImageNet in 420 GPU hours, reducing the total time by more than 34%.
****

SESS: Self-Ensembling Semi-Supervised 3D Object Detection
Na Zhao, Tat-Seng Chua, Gim Hee Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11079-11087
The performance of existing point cloud-based 3D object detection methods heavily relies on large-scale high-quality 3D annotations. However, such annotations are often tedious and expensive to collect. Semi-supervised learning is a good alternative to mitigate the data annotation issue, but has remained largely unexplored in 3D object detection. Inspired by the recent success of self-ensembling technique in semi-supervised image classification task, we propose SESS, a self-ensembling semi-supervised 3D object detection framework. Specifically, we design a thorough perturbation scheme to enhance generalization of the network on unlabeled and new unseen data. Furthermore, we propose three consistency losses to enforce the consistency between two sets of predicted 3D object proposals, to facilitate the learning of structure and semantic invariances of objects. Extensive experiments conducted on SUN RGB-D and ScanNet datasets demonstrate the effectiveness of SESS in both inductive and transductive semi-supervised 3D object detection. Our SESS achieves competitive performance compared to the state-of-the-art fully-supervised method by using only 50% labeled data. Our code is available at https://github.com/Na-Z/sess.
****

Panoptic-DeepLab: A Simple, Strong, and Fast Baseline for Bottom-Up Panoptic Segmentation
Bowen Cheng, Maxwell D. Collins, Yukun Zhu, Ting Liu, Thomas S. Huang, Hartwig Adam, Liang-Chieh Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12475-12485
In this work, we introduce Panoptic-DeepLab, a simple, strong, and fast system for panoptic segmentation, aiming to establish a solid baseline for bottom-up methods that can achieve comparable performance of two-stage methods while yielding fast inference speed. In particular, Panoptic-DeepLab adopts the dual-ASPP and dual-decoder structures specific to semantic, and instance segmentation, respect

ively. The semantic segmentation branch is the same as the typical design of any semantic segmentation model (e.g., DeepLab), while the instance segmentation branch is class-agnostic, involving a simple instance center regression. As a result, our single Panoptic-DeepLab simultaneously ranks first at all three Cityscapes benchmarks, setting the new state-of-art of 84.2% mIoU, 39.0% AP, and 65.5% PQ on test set. Additionally, equipped with MobileNetV3, Panoptic-DeepLab runs nearly in real-time with a single 1025x2049 image (15.8 frames per second), while achieving a competitive performance on Cityscapes (54.1 PQ% on test set). On Mapillary Vistas test set, our ensemble of six models attains 42.7% PQ, outperforming the challenge winner in 2018 by a healthy margin of 1.5%. Finally, our Panoptic-DeepLab also performs on par with several top-down approaches on the challenging COCO dataset. For the first time, we demonstrate a bottom-up approach could deliver state-of-the-art results on panoptic segmentation.

****

Spatio-Temporal Graph for Video Captioning With Knowledge Distillation

Boxiao Pan, Haoye Cai, De-An Huang, Kuan-Hui Lee, Adrien Gaidon, Ehsan Adeli, Juan Carlos Niebles; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10870-10879

Video captioning is a challenging task that requires a deep understanding of visual scenes. State-of-the-art methods generate captions using either scene-level or object-level information but without explicitly modeling object interactions. Thus, they often fail to make visually grounded predictions, and are sensitive to spurious correlations. In this paper, we propose a novel spatio-temporal graph model for video captioning that exploits object interactions in space and time. Our model builds interpretable links and is able to provide explicit visual grounding. To avoid unstable performance caused by the variable number of objects, we further propose an object-aware knowledge distillation mechanism, in which local object information is used to regularize global scene features. We demonstrate the efficacy of our approach through extensive experiments on two benchmarks, showing our approach yields competitive performance with interpretable predictions.

****

ACNe: Attentive Context Normalization for Robust Permutation-Equivariant Learning

Weiwei Sun, Wei Jiang, Eduard Trulls, Andrea Tagliasacchi, Kwang Moo Yi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11286-11295

Many problems in computer vision require dealing with sparse, unordered data in the form of point clouds. Permutation-equivariant networks have become a popular solution - they operate on individual data points with simple perceptrons and extract contextual information with global pooling. This can be achieved with a simple normalization of the feature maps, a global operation that is unaffected by the order. In this paper, we propose Attentive Context Normalization (ACN), a simple yet effective technique to build permutation-equivariant networks robust to outliers. Specifically, we show how to normalize the feature maps with weights that are estimated within the network, excluding outliers from this normalization. We use this mechanism to leverage two types of attention: local and global - by combining them, our method is able to find the essential data points in high-dimensional space in order to solve a given task. We demonstrate through extensive experiments that our approach, which we call Attentive Context Networks (ACNe), provides a significant leap in performance compared to the state-of-the-art on camera pose estimation, robust fitting, and point cloud classification under noise and outliers. Source code: https://github.com/vcg-uvic/acne.

****

ViBE: Dressing for Diverse Body Shapes

Wei-Lin Hsiao, Kristen Grauman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11059-11069

Body shape plays an important role in determining what garments will best suit a given person, yet today's clothing recommendation methods take a "one shape fits all" approach. These body-agnostic vision methods and datasets are a barrier t

o inclusion, ill-equipped to provide good suggestions for diverse body shapes. We introduce ViBE, a VIsual Body-aware Embedding that captures clothing's affinity with different body shapes. Given an image of a person, the proposed embedding identifies garments that will flatter her specific body shape. We show how to learn the embedding from an online catalog displaying fashion models of various shapes and sizes wearing the products, and we devise a method to explain the algorithm's suggestions for well-fitting garments. We apply our approach to a dataset of diverse subjects, and demonstrate its strong advantages over status quo body-agnostic recommendation, both according to automated metrics and human opinion.

****

Density-Based Clustering for 3D Object Detection in Point Clouds

Syeda Mariam Ahmed, Chee Meng Chew; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10608-10617

Current 3D detection networks either rely on 2D object proposals or try to directly predict bounding box parameters from each point in a scene. While former methods are dependent on performance of 2D detectors, latter approaches are challenging due to the sparsity and occlusion in point clouds, making it difficult to regress accurate parameters. In this work, we introduce a novel approach for 3D object detection that is significant in two main aspects: a) cascaded modular approach that focuses the receptive field of each module on specific points in the point cloud, for improved feature learning and b) a class agnostic instance segmentation module that is initiated using unsupervised clustering. The objective of a cascaded approach is to sequentially minimize the number of points running through the network. While three different modules perform the tasks of background-foreground segmentation, class agnostic instance segmentation and object detection, through individually trained point based networks. We also evaluate bayesian uncertainty in modules, demonstrating the over all level of confidence in our prediction results. Performance of the network is evaluated on the SUN RGB-D benchmark dataset, that demonstrates an improvement as compared to state-of-the-art methods.

****

Diverse Image Generation via Self-Conditioned GANs

Steven Liu, Tongzhou Wang, David Bau, Jun-Yan Zhu, Antonio Torralba; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14286-14295

We introduce a simple but effective unsupervised method for generating diverse images. We train a class-conditional GAN model without using manually annotated class labels. Instead, our model is conditional on labels automatically derived from clustering in the discriminator's feature space. Our clustering step automatically discovers diverse modes, and explicitly requires the generator to cover them. Experiments on standard mode collapse benchmarks show that our method outperforms several competing methods when addressing mode collapse. Our method also performs well on large-scale datasets such as ImageNet and Places365, improving both diversity and standard metrics (e.g., Frechet Inception Distance), compared to previous methods.

****

A Certifiably Globally Optimal Solution to Generalized Essential Matrix Estimation

Ji Zhao, Wanting Xu, Laurent Kneip; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12034-12043

We present a convex optimization approach for generalized essential matrix (GEM) estimation. The six-point minimal solver for the GEM has poor numerical stability and applies only for a minimal number of points. Existing non-minimal solvers for GEM estimation rely on either local optimization or relinearization techniques, which impedes high accuracy in common scenarios. Our proposed non-minimal solver minimizes the sum of squared residuals by reformulating the problem as a quadratically constrained quadratic program. The globally optimal solution is thus obtained by a semidefinite relaxation. The algorithm retrieves certifiably globally optimal solutions to the original non-convex problem in polynomial time. W

e also provide the necessary and sufficient conditions to recover the optimal GE
M from the relaxed problems. The improved performance is demonstrated over exper
iments on both synthetic and real multi-camera systems.
****

Video Panoptic Segmentation
Dahun Kim, Sanghyun Woo, Joon-Young Lee, In So Kweon; Proceedings of the IEEE
/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 985
9-9868
Panoptic segmentation has become a new standard of visual recognition task by un
ifying previous semantic segmentation and instance segmentation tasks in concert
. In this paper, we propose and explore a new video extension of this task, call
ed video panoptic segmentation. The task requires generating consistent panoptic
 segmentation as well as an association of instance ids across video frames. To
invigorate research on this new task, we present two types of video panoptic dat
asets. The first is a re-organization of the synthetic VIPER dataset into the vi
deo panoptic format to exploit its large-scale pixel annotations. The second is
a temporal extension on the Cityscapes val. set, by providing new video panoptic
 annotations (Cityscapes-VPS). Moreover, we propose a novel video panoptic segme
ntation network (VPSNet) which jointly predicts object classes, bounding boxes,
masks, instance id tracking, and semantic segmentation in video frames. To provi
de appropriate metrics for this task, we propose a video panoptic quality (VPQ)
metric and evaluate our method and several other baselines. Experimental results
 demonstrate the effectiveness of the presented two datasets. We achieve state-o
f-the-art results in image PQ on Cityscapes and also in VPQ on Cityscapes-VPS an
d VIPER datasets.
****

Structured Multi-Hashing for Model Compression
Elad Eban, Yair Movshovitz-Attias, Hao Wu, Mark Sandler, Andrew Poon, Yerla
n Idelbayev, Miguel A. Carreira-Perpinan; Proceedings of the IEEE/CVF Conferenc
e on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11903-11912
Despite the success of deep neural networks (DNNs), state-of-the-art models are
too large to deploy on low-resource devices or common server configurations in w
hich multiple models are held in memory. Model compression methods address this
limitation by reducing the memory footprint, latency, or energy consumption of a
 model with minimal impact on accuracy. We focus on the task of reducing the num
ber of learnable variables in the model. In this work we combine ideas from weig
ht hashing and dimensionality reductions resulting in a simple and powerful stru
ctured multi-hashing method based on matrix products that allows direct control
of model size of any deep network and is trained end-to-end. We demonstrate the
strength of our approach by compressing models from the ResNet, EfficientNet, an
d MobileNet architecture families. Our method allows us to drastically decrease
the number of variables while maintaining high accuracy. For instance, by applyi
ng our approach to EfficentNet-B4 (16M parameters) we reduce it to the size of B
0 (5M parameters), while gaining over 3% in accuracy over B0 baseline. On the co
mmonly used benchmark CIFAR10 we reduce the ResNet32 model by 75% with no loss i
n quality, and are able to do a 10x compression while still achieving above 90%
accuracy.
****

Maintaining Discrimination and Fairness in Class Incremental Learning
Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, Shu-Tao Xia; Proceedings of the
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp.
 13208-13217
Deep neural networks (DNNs) have been applied in class incremental learning, whi
ch aims to solve common real-world problems of learning new classes continually.
 One drawback of standard DNNs is that they are prone to catastrophic forgetting
. Knowledge distillation (KD) is a commonly used technique to alleviate this pro
blem. In this paper, we demonstrate it can indeed help the model to output more
discriminative results within old classes. However, it cannot alleviate the prob
lem that the model tends to classify objects into new classes, causing the posit
ive effect of KD to be hidden and limited. We observed that an important factor

causing catastrophic forgetting is that the weights in the last fully connected (FC) layer are highly biased in class incremental learning. In this paper, we propose a simple and effective solution motivated by the aforementioned observations to address catastrophic forgetting. Firstly, we utilize KD to maintain the discrimination within old classes. Then, to further maintain the fairness between old classes and new classes, we propose Weight Aligning (WA) that corrects the biased weights in the FC layer after normal training process. Unlike previous work, WA does not require any extra parameters or a validation set in advance, as it utilizes the information provided by the biased weights themselves. The proposed method is evaluated on ImageNet-1000, ImageNet-100, and CIFAR-100 under various settings. Experimental results show that the proposed method can effectively alleviate catastrophic forgetting and significantly outperform state-of-the-art methods.

****

ZeroQ: A Novel Zero Shot Quantization Framework

Yaohui Cai, Zhewei Yao, Zhen Dong, Amir Gholami, Michael W. Mahoney, Kurt Keutzer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13169-13178

Quantization is a promising approach for reducing the inference time and memory footprint of neural networks. However, most existing quantization methods require access to the original training dataset for retraining during quantization. This is often not possible for applications with sensitive or proprietary data, e.g., due to privacy and security concerns. Existing zero-shot quantization methods use different heuristics to address this, but they result in poor performance, especially when quantizing to ultra-low precision. Here, we propose \OURS, a novel zero-shot quantization framework to address this. \OURS enables mixed-precision quantization without any access to the training or validation data. This is achieved by optimizing for a Distilled Dataset, which is engineered to match the statistics of batch normalization across different layers of the network. \OURS supports both uniform and mixed-precision quantization. For the latter, we introduce a novel Pareto frontier based method to automatically determine the mixed-precision bit setting for all layers, with no manual search involved. We extensively test our proposed method on a diverse set of models, including ResNet18/50/152, MobileNetV2, ShuffleNet, SqueezeNext, and InceptionV3 on ImageNet, as well as RetinaNet-ResNet50 on the Microsoft COCO dataset. In particular, we show that \OURS can achieve 1.71% higher accuracy on MobileNetV2, as compared to the recently proposed DFQ [??] method. Importantly, \OURS has a very low computational overhead, and it can finish the entire quantization process in less than 30s (0.5% of one epoch training time of ResNet50 on ImageNet). We have open-sourced the \OURS framework(https://github.com/amirgholami/ZeroQ).

****

Learning Visual Motion Segmentation Using Event Surfaces

Anton Mitrokhin, Zhiyuan Hua, Cornelia Fermuller, Yiannis Aloimonos; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14414-14423

Event-based cameras have been designed for scene motion perception - their high temporal resolution and spatial data sparsity converts the scene into a volume of boundary trajectories and allows to track and analyze the evolution of the scene in time. Analyzing this data is computationally expensive, and there is substantial lack of theory on dense-in-time object motion to guide the development of new algorithms; hence, many works resort to a simple solution of discretizing the event stream and converting it to classical pixel maps, which allows for application of conventional image processing methods. In this work we present a Graph Convolutional neural network for the task of scene motion segmentation by a moving camera. We convert the event stream into a 3D graph in (x,y,t) space and keep per-event temporal information. The difficulty of the task stems from the fact that unlike in metric space, the shape of an object in (x,y,t) space depends on its motion and is not the same across the dataset. We discuss properties of of the event data with respect to this 3D recognition problem, and show that our Graph Convolutional architecture is superior to PointNet++. We evaluate our metho

d on the state of the art event-based motion segmentation dataset - EV-IMO and p erform comparisons to a frame-based method proposed by its authors. Our ablation studies show that increasing the event slice width improves the accuracy, and h ow subsampling and edge configurations affect the network performance.
****

Orthogonal Convolutional Neural Networks
Jiayun Wang, Yubei Chen, Rudrasis Chakraborty, Stella X. Yu; Proceedings of t he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11505-11515
Deep convolutional neural networks are hindered by training instability and feat ure redundancy towards further performance improvement. A promising solution is to impose orthogonality on convolutional filters. We develop an efficient approa ch to impose filter orthogonality on a convolutional layer based on the doubly b lock-Toeplitz matrix representation of the convolutional kernel, instead of the common kernel orthogonality approach, which we show is only necessary but not su fficient for ensuring orthogonal convolutions. Our proposed orthogonal convoluti on requires no additional parameters and little computational overhead. It consi stently outperforms the kernel orthogonality alternative on a wide range of task s such as image classification and inpainting under supervised, semi-supervised and unsupervised settings. It learns more diverse and expressive features with b etter training stability, robustness, and generalization. Our code is publicly a vailable.
****

Just Go With the Flow: Self-Supervised Scene Flow Estimation
Himangi Mittal, Brian Okorn, David Held; Proceedings of the IEEE/CVF Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11177-11185
When interacting with highly dynamic environments, scene flow allows autonomous systems to reason about the non-rigid motion of multiple independent objects. Th is is of particular interest in the field of autonomous driving, in which many c ars, people, bicycles, and other objects need to be accurately tracked. Current state-of-the-art methods require annotated scene flow data from autonomous drivi ng scenes to train scene flow networks with supervised learning. As an alternati ve, we present a method of training scene flow that uses two self-supervised los ses, based on nearest neighbors and cycle consistency. These self-supervised los ses allow us to train our method on large unlabeled autonomous driving datasets; the resulting method matches current state-of-the-art supervised performance us ing no real world annotations and exceeds state-of-the-art performance when comb ining our self-supervised approach with supervised learning on a smaller labeled dataset.
****

Set-Constrained Viterbi for Set-Supervised Action Segmentation
Jun Li, Sinisa Todorovic; Proceedings of the IEEE/CVF Conference on Computer Vi sion and Pattern Recognition (CVPR), 2020, pp. 10820-10829
This paper is about weakly supervised action segmentation, where the ground trut h specifies only a set of actions present in a training video, but not their tru e temporal ordering. Prior work typically uses a classifier that independently l abels video frames for generating the pseudo ground truth, and multiple instance learning for training the classifier. We extend this framework by specifying an HMM, which accounts for co-occurrences of action classes and their temporal len gths, and by explicitly training the HMM on a Viterbi-based loss. Our first cont ribution is the formulation of a new set-constrained Viterbi algorithm (SCV). Gi ven a video, the SCV generates the MAP action segmentation that satisfies the gr ound truth. This prediction is used as a framewise pseudo ground truth in our HM M training. Our second contribution in training is a new regularization of featu re affinities between training videos that share the same action classes. Evalua tion on action segmentation and alignment on the Breakfast, MPII Cooking2, Holly wood Extended datasets demonstrates our significant performance improvement for the two tasks over prior work.
****

Fast Sparse ConvNets

Erich Elsen, Marat Dukhan, Trevor Gale, Karen Simonyan; Proceedings of the IE EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1 4629-14638
Historically, the pursuit of efficient inference has been one of the driving for ces behind the research into new deep learning architectures and building blocks . Some of the recent examples include: the squeeze-and-excitation module, depthw ise separable convolutions in Xception, and the inverted bottleneck in MobileNet v2. Notably, in all of these cases, the resulting building blocks enabled not o nly higher efficiency, but also higher accuracy, and found wide adoption in the field. In this work, we further expand the arsenal of efficient building blocks for neural network architectures; but instead of combining standard primitives (such as convolution), we advocate for the replacement of these dense primitives with their sparse counterparts. While the idea of using sparsity to decrease the parameter count is not new, the conventional wisdom is that this reduction in t heoretical FLOPs does not translate into real-world efficiency gains. We aim to correct this misconception by introducing a family of efficient sparse kernels f or several hardware platforms, which we plan to open source for the benefit of t he community. Equipped with our efficient implementation of sparse primitives, w e show that sparse versions of MobileNet v1 and MobileNet v2 architectures subst antially outperform strong dense baselines on the efficiency-accuracy curve. On Snapdragon 835 our sparse networks outperform their dense equivalents by 1.3 - 2 .4x - equivalent to approximately one entire generation of improvement. We hope that our findings will facilitate wider adoption of sparsity as a tool for creat ing efficient and accurate deep learning architectures.
****

Learning a Weakly-Supervised Video Actor-Action Segmentation Model With a Wise S election
Jie Chen, Zhiheng Li, Jiebo Luo, Chenliang Xu; Proceedings of the IEEE/CVF Co nference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9901-9911
We address weakly-supervised video actor-action segmentation (VAAS), which exten ds general video object segmentation (VOS) to additionally consider action label s of the actors. The most successful methods on VOS synthesize a pool of pseudo- annotations (PAs) and then refine them iteratively. However, they face challenge s as to how to select from a massive amount of PAs high-quality ones, how to set an appropriate stop condition for weakly-supervised training, and how to initia lize PAs pertaining to VAAS. To overcome these challenges, we propose a general Weakly-Supervised framework with a Wise Selection of training samples and model evaluation criterion (WS^2). Instead of blindly trusting quality-inconsistent PA s, WS^2 employs a learning-based selection to select effective PAs and a novel r egion integrity criterion as a stopping condition for weakly-supervised training . In addition, a 3D-Conv GCAM is devised to adapt to the VAAS task. Extensive ex periments show that WS^2 achieves state-of-the-art performance on both weakly-su pervised VOS and VAAS tasks and is on par with the best fully-supervised method on VAAS.
****

Gradually Vanishing Bridge for Adversarial Domain Adaptation
Shuhao Cui, Shuhui Wang, Junbao Zhuo, Chi Su, Qingming Huang, Qi Tian; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2020, pp. 12455-12464
In unsupervised domain adaptation, rich domain-specific characteristics bring gr eat challenge to learn domain-invariant representations. However, domain discrep ancy is considered to be directly minimized in existing solutions, which is diff icult to achieve in practice. Some methods alleviate the difficulty by explicitl y modeling domain-invariant and domain-specific parts in the representations, bu t the adverse influence of the explicit construction lies in the residual domain -specific characteristics in the constructed domain-invariant representations. I n this paper, we equip adversarial domain adaptation with Gradually Vanishing Br idge (GVB) mechanism on both generator and discriminator. On the generator, GVB could not only reduce the overall transfer difficulty, but also reduce the influ ence of the residual domain-specific characteristics in domain-invariant represe

ntations. On the discriminator, GVB contributes to enhance the discriminating ab ility, and balance the adversarial training process. Experiments on three challe nging datasets show that our GVB methods outperform strong competitors, and coop erate well with other adversarial methods. The code is available at https://gith ub.com/cuishuhao/GVB.
****

Deep Degradation Prior for Low-Quality Image Classification
Yang Wang, Yang Cao, Zheng-Jun Zha, Jing Zhang, Zhiwei Xiong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020 , pp. 11049-11058
State-of-the-art image classification algorithms building upon convolutional neu ral networks (CNNs) are commonly trained on large annotated datasets of high-qua lity images. When applied to low-quality images, they will suffer a significant degradation in performance, since the structural and statistical properties of p ixels in the neighborhood are obstructed by image degradation. To address this p roblem, this paper proposes a novel deep degradation prior for low-quality image classification. It is based on statistical observations that, in the deep repre sentation space, image patches with structural similarity have uniform distribut ion even if they come from different images, and the distributions of correspond ing patches in low- and high-quality images have uniform margins under the same degradation condition. Therefore, we propose a feature de-drifting module (FDM) to learn the mapping relationship between deep representations of low- and high- quality images, and leverage it as a deep degradation prior (DDP) for low-quali ty image classification. Since the statistical properties are independent to ima ge content, deep degradation prior can be learned on a training set of limited i mages without supervision of semantic labels and served in a form of "plugging-i n" module of the existing classification networks to improve their performance o n degraded images. Evaluations on the benchmark dataset ImageNet-C demonstrate t hat our proposed DDP can improve the accuracy of the pre-trained network model b y more than 20% under various degradation conditions. Even under the extreme set ting that only 10 images from CUB-C dataset are used for the training of DDP, ou r method improves the accuracy of VGG16 on ImageNet-C from 37% to 55%.
****

Visual-Textual Capsule Routing for Text-Based Video Segmentation
Bruce McIntosh, Kevin Duarte, Yogesh S Rawat, Mubarak Shah; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, p p. 9942-9951
Joint understanding of vision and natural language is a challenging problem with a wide range of applications in artificial intelligence. In this work, we focus on integration of video and text for the task of actor and action video segment ation from a sentence. We propose a capsule-based approach which performs pixel- level localization based on a natural language query describing the actor of int erest. We encode both the video and textual input in the form of capsules, which provide a more effective representation in comparison with standard convolution based features. Our novel visual-textual routing mechanism allows for the fusio n of video and text capsules to successfully localize the actor and action. The existing works on actor-action localization are mainly focused on localization i n a single frame instead of the full video. Different from existing works, we pr opose to perform the localization on all frames of the video. To validate the po tential of the proposed network for actor and action video localization, we exte nd an existing actor-action dataset (A2D) with annotations for all the frames. T he experimental evaluation demonstrates the effectiveness of our capsule network for text selective actor and action localization in videos. The proposed method also improves upon the performance of the existing state-of-the art works on si ngle frame-based localization.
****

Towards Inheritable Models for Open-Set Domain Adaptation
Jogendra Nath Kundu, Naveen Venkat, Ambareesh Revanur, Rahul M V, R. Venkate sh Babu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern R ecognition (CVPR), 2020, pp. 12376-12385

There has been a tremendous progress in Domain Adaptation (DA) for visual recogn ition tasks. Particularly, open-set DA has gained considerable attention wherein the target domain contains additional unseen categories. Existing open-set DA a pproaches demand access to a labeled source dataset along with unlabeled target instances. However, this reliance on co-existing source and target data is highl y impractical in scenarios where data-sharing is restricted due to its proprieta ry nature or privacy concerns. Addressing this, we introduce a practical DA para digm where a source-trained model is used to facilitate adaptation in the absenc e of the source dataset in future. To this end, we formalize knowledge inheritab ility as a novel concept and propose a simple yet effective solution to realize inheritable models suitable for the above practical paradigm. Further, we presen t an objective way to quantify inheritability to enable the selection of the mos t suitable source model for a given target domain, even in the absence of the so urce data. We provide theoretical insights followed by a thorough empirical eval uation demonstrating state-of-the-art open-set domain adaptation performance.
****

Multi-Task Collaborative Network for Joint Referring Expression Comprehension an d Segmentation

Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, R ongrong Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Patter n Recognition (CVPR), 2020, pp. 10034-10043

Referring expression comprehension (REC) and segmentation (RES) are two highly-r elated tasks, which both aim at identifying the referent according to a natural language expression. In this paper, we propose a novel Multi-task Collaborative Network (MCN) to achieve a joint learning of REC and RES for the first time. In MCN, RES can help REC to achieve better language-vision alignment, while REC can help RES to better locate the referent. In addition, we address a key challenge in this multi-task setup, i.e., the prediction conflict, with two innovative de signs namely, Consistency Energy Maximization (CEM) and Adaptive Soft Non-Locate d Suppression (ASNLS). Specifically, CEM enables REC and RES to focus on similar visual regions by maximizing the consistency energy between two tasks. ASNLS su presses the response of unrelated regions in RES based on the prediction of REC. To validate our model, we conduct extensive experiments on three benchmark data sets of REC and RES, i.e., RefCOCO, RefCOCO+ and RefCOCOg. The experimental resu lts report the significant performance gains of MCN over all existing methods, i .e., up to +7.13% for REC and +11.50% for RES over SOTA, which well confirm the validity of our model for joint REC and RES learning.
****

Where, What, Whether: Multi-Modal Learning Meets Pedestrian Detection

Yan Luo, Chongyang Zhang, Muming Zhao, Hao Zhou, Jun Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp . 14065-14073

Pedestrian detection benefits greatly from deep convolutional neural networks (C NNs). However, it is inherently hard for CNNs to handle situations in the presen ce of occlusion and scale variation. In this paper, we propose W^3Net, which att empts to address above challenges by decomposing the pedestrian detection task i nto Where, What and Whether problem directing against pedestrian localization, s cale prediction and classification correspondingly. Specifically, for a pedestri an instance, we formulate its feature by three steps. i) We generate a bird view map, which is naturally free from occlusion issues, and scan all points on it t o look for suitable locations for each pedestrian instance. ii) Instead of utili zing pre-fixed anchors, we model the interdependency between depth and scale aim ing at generating depth-guided scales at different locations for better matching instances of different sizes. iii) We learn a latent vector shared by both visu al and corpus space, by which false positives with similar vertical structure bu t lacking human partial features would be filtered out. We achieve state-of-the- art results on widely used datasets (Citypersons and Caltech). In particular. wh en evaluating on heavy occlusion subset, our results reduce MR^ -2 from 49.3% t o 18.7% on Citypersons, and from 45.18% to 28.33% on Caltech.
****

Learning Depth-Guided Convolutions for Monocular 3D Object Detection
Mingyu Ding, Yuqi Huo, Hongwei Yi, Zhe Wang, Jianping Shi, Zhiwu Lu, Ping Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11672-11681
3D object detection from a single image without LiDAR is a challenging task due to the lack of accurate depth information. Conventional 2D convolutions are unsuitable for this task because they fail to capture local object and its scale information, which are vital for 3D object detection. To better represent 3D structure, prior arts typically transform depth maps estimated from 2D images into a pseudo-LiDAR representation, and then apply existing 3D point-cloud based object detectors. However, their results depend heavily on the accuracy of the estimated depth maps, resulting in suboptimal performance. In this work, instead of using pseudo-LiDAR representation, we improve the fundamental 2D fully convolutions by proposing a new local convolutional network (LCN), termed Depth-guided Dynamic-Depthwise-Dilated LCN (D4LCN), where the filters and their receptive fields can be automatically learned from image-based depth maps, making different pixels of different images have different filters. D4LCN overcomes the limitation of conventional 2D convolutions and narrows the gap between image representation and 3D point cloud representation. Extensive experiments show that D^4LCN outperforms existing works by large margins. For example, the relative improvement of D4LCN against the state-of-the-art on KITTI is 9.1% in the moderate setting. D4LCN ranks 1st on KITTI monocular 3D object detection benchmark at the time of submission (car, December 2019). The code is available at https://github.com/dingmyu/D4LCN
****
Your Local GAN: Designing Two Dimensional Local Attention Mechanisms for Generative Models
Giannis Daras, Augustus Odena, Han Zhang, Alexandros G. Dimakis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14531-14539
We introduce a new local sparse attention layer that preserves two-dimensional geometry and locality. We show that by just replacing the dense attention layer of SAGAN with our construction, we obtain very significant FID, Inception score and pure visual improvements. FID score is improved from 18.65 to 15.94 on ImageNet, keeping all other parameters the same. The sparse attention patterns that we propose for our new layer are designed using a novel information theoretic criterion that uses information flow graphs. We also present a novel way to invert Generative Adversarial Networks with attention. Our method uses the attention layer of the discriminator to create an innovative loss function. This allows us to visualize the newly introduced attention heads and show that they indeed capture interesting aspects of two-dimensional geometry of real images.
****
Context Aware Graph Convolution for Skeleton-Based Action Recognition
Xikun Zhang, Chang Xu, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14333-14342
Graph convolutional models have gained impressive successes on skeleton based human action recognition task. As graph convolution is a local operation, it cannot fully investigate non-local joints that could be vital to recognizing the action. For example, actions like typing and clapping request the cooperation of two hands, which are distant from each other in a human skeleton graph. Multiple graph convolutional layers thus tend to be stacked together to increase receptive field, which brings in computational inefficiency and optimization difficulty. But there is still no guarantee that distant joints (e.g. two hands) can be well integrated. In this paper, we propose a context aware graph convolutional network (CA-GCN). Besides the computation of localized graph convolution, CA-GCN considers a context term for each vertex by integrating information of all other vertices. Long range dependencies among joints are thus naturally integrated in context information, which then eliminates the need of stacking multiple layers to enlarge receptive field and greatly simplifies the network. Moreover, we further propose an advanced CA-GCN, in which asymmetric relevance measurement and higher

level representation are utilized to compute context information for more flexibility and better performance. Besides the joint features, our CA-GCN could also be extended to handle graphs with edge (limb) features. Extensive experiments on two real-world datasets demonstrate the importance of context information and the effectiveness of the proposed CA-GCN in skeleton based action recognition.
****

Probabilistic Video Prediction From Noisy Data With a Posterior Confidence
Yunbo Wang, Jiajun Wu, Mingsheng Long, Joshua B. Tenenbaum; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10830-10839
We study a new research problem of probabilistic future frames prediction from a sequence of noisy inputs, which is useful because it is difficult to guarantee the quality of input frames in practical spatiotemporal prediction applications. It is also challenging because it involves two levels of uncertainty: the perceptual uncertainty from noisy observations and the dynamics uncertainty in forward modeling. In this paper, we propose to tackle this problem with an end-to-end trainable model named Bayesian Predictive Network (BP-Net). Unlike previous work in stochastic video prediction that assumes spatiotemporal coherence and therefore fails to deal with perceptual uncertainty, BP-Net models both levels of uncertainty in an integrated framework. Furthermore, unlike previous work that can only provide unsorted estimations of future frames, BP-Net leverages a differentiable sequential importance sampling (SIS) approach to make future predictions based on the inference of underlying physical states, thereby providing sorted prediction candidates in accordance with the SIS importance weights, i.e., the confidences. Our experiment results demonstrate that BP-Net remarkably outperforms existing approaches on predicting future frames from noisy data.
****

Generalizing Hand Segmentation in Egocentric Videos With Uncertainty-Guided Model Adaptation
Minjie Cai, Feng Lu, Yoichi Sato; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14392-14401
Although the performance of hand segmentation in egocentric videos has been significantly improved by using CNNs, it still remains a challenging issue to generalize the trained models to new domains, e.g., unseen environments. In this work, we solve the hand segmentation generalization problem without requiring segmentation labels in the target domain. To this end, we propose a Bayesian CNN-based model adaptation framework for hand segmentation, which introduces and considers two key factors: 1) prediction uncertainty when the model is applied in a new domain and 2) common information about hand shapes shared across domains. Consequently, we propose an iterative self-training method for hand segmentation in the new domain, which is guided by the model uncertainty estimated by a Bayesian CNN. We further use an adversarial component in our framework to utilize shared information about hand shapes to constrain the model adaptation process. Experiments on multiple egocentric datasets show that the proposed method significantly improves the generalization performance of hand segmentation.
****

Revisiting Pose-Normalization for Fine-Grained Few-Shot Recognition
Luming Tang, Davis Wertheimer, Bharath Hariharan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14352-14361
Few-shot, fine-grained classification requires a model to learn subtle, fine-grained distinctions between different classes (e.g., birds) based on a few images alone. This requires a remarkable degree of invariance to pose, articulation and background. A solution is to use pose-normalized representations: first localize semantic parts in each image, and then describe images by characterizing the appearance of each part. While such representations are out of favor for fully supervised classification, we show that they are extremely effective for few-shot fine-grained classification. With a minimal increase in model capacity, pose normalization improves accuracy between 10 and 20 percentage points for shallow and deep architectures, generalizes better to new domains, and is effective for mul

tiple few-shot algorithms and network backbones. Code is available at https://github.com/Tsingularity/PoseNorm_Fewshot.
****

Weakly-Supervised Salient Object Detection via Scribble Annotations
Jing Zhang, Xin Yu, Aixuan Li, Peipei Song, Bowen Liu, Yuchao Dai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12546-12555
Compared with laborious pixel-wise dense labeling, it is much easier to label data by scribbles, which only costs 1 2 seconds to label one image. However, using scribble labels to learn salient object detection has not been explored. In this paper, we propose a weakly-supervised salient object detection model to learn saliency from such annotations. In doing so, we first relabel an existing large-scale salient object detection dataset with scribbles, namely S-DUTS dataset. Since object structure and detail information is not identified by scribbles, directly training with scribble labels will lead to saliency maps of poor boundary localization. To mitigate this problem, we propose an auxiliary edge detection task to localize object edges explicitly, and a gated structure-aware loss to place constraints on the scope of structure to be recovered. Moreover, we design a scribble boosting scheme to iteratively consolidate our scribble annotations, which are then employed as supervision to learn high-quality saliency maps. As existing saliency evaluation metrics neglect to measure structure alignment of the predictions, the saliency map ranking may not comply with human perception. We present a new metric, termed saliency structure measure, as a complementary metric to evaluate sharpness of the prediction. Extensive experiments on six benchmark datasets demonstrate that our method not only outperforms existing weakly-supervised/unsupervised methods, but also is on par with several fully-supervised state-of-the-art models (Our code and data is publicly available at: https://github.com/JingZhang617/Scribble_Saliency).
****

Correspondence Networks With Adaptive Neighbourhood Consensus
Shuda Li, Kai Han, Theo W. Costain, Henry Howard-Jenkins, Victor Prisacariu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10196-10205
In this paper, we tackle the task of establishing dense visual correspondences between images containing objects of the same category. This is a challenging task due to large intra-class variations and a lack of dense pixel level annotations. We propose a convolutional neural network architecture, called adaptive neighbourhood consensus network (ANC-Net), that can be trained end-to-end with sparse key-point annotations, to handle this challenge. At the core of ANC-Net is our proposed non-isotropic 4D convolution kernel, which forms the building block for the adaptive neighbourhood consensus module for robust matching. We also introduce a simple and efficient multi-scale self-similarity module in ANC-Net to make the learned feature robust to intra-class variations. Furthermore, we propose a novel orthogonal loss that can enforce the one-to-one matching constraint. We thoroughly evaluate the effectiveness of our method on various benchmarks, where it substantially outperforms state-of-the-art methods.
****

Interactive Object Segmentation With Inside-Outside Guidance
Shiyin Zhang, Jun Hao Liew, Yunchao Wei, Shikui Wei, Yao Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12234-12244
This paper explores how to harvest precise object segmentation masks while minimizing the human interaction cost. To achieve this, we propose an Inside-Outside Guidance (IOG) approach in this work. Concretely, we leverage an inside point that is clicked near the object center and two outside points at the symmetrical corner locations (top-left and bottom-right or top-right and bottom-left) of a tight bounding box that encloses the target object. This results in a total of one foreground click and four background clicks for segmentation. The advantages of our IOG is four-fold: 1) the two outside points can help to remove distractions from other objects or background; 2) the inside point can help to eliminate the

unrelated regions inside the bounding box; 3) the inside and outside points are easily identified, reducing the confusion raised by the state-of-the-art DEXTR in labeling some extreme samples; 4) our approach naturally supports additional clicks annotations for further correction. Despite its simplicity, our IOG not only achieves state-of-the-art performance on several popular benchmarks, but also demonstrates strong generalization capability across different domains such as street scenes, aerial imagery and medical images, without fine-tuning. In addition, we also propose a simple two-stage solution that enables our IOG to produce high quality instance segmentation masks from existing datasets with off-the-shelf bounding boxes such as ImageNet and Open Images, demonstrating the superiority of our IOG as an annotation tool.
****

GraspNet-1Billion: A Large-Scale Benchmark for General Object Grasping
Hao-Shu Fang, Chenxi Wang, Minghao Gou, Cewu Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11444-11453
Object grasping is critical for many applications, which is also a challenging computer vision problem. However, for cluttered scene, current researches suffer from the problems of insufficient training data and the lacking of evaluation benchmarks. In this work, we contribute a large-scale grasp pose detection dataset with a unified evaluation system. Our dataset contains 97,280 RGB-D image with over one billion grasp poses. Meanwhile, our evaluation system directly reports whether a grasping is successful by analytic computation, which is able to evaluate any kind of grasp poses without exhaustively labeling ground-truth. In addition, we propose an end-to-end grasp pose prediction network given point cloud inputs, where we learn approaching direction and operation parameters in a decoupled manner. A novel grasp affinity field is also designed to improve the grasping robustness. We conduct extensive experiments to show that our dataset and evaluation system can align well with real-world experiments and our proposed network achieves the state-of-the-art performance. Our dataset, source code and models are publicly available at www.graspnet.net.
****

Meshed-Memory Transformer for Image Captioning
Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, Rita Cucchiara; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10578-10587
Transformer-based architectures represent the state of the art in sequence modeling tasks like machine translation and language understanding. Their applicability to multi-modal contexts like image captioning, however, is still largely under-explored. With the aim of filling this gap, we present M2 - a Meshed Transformer with Memory for Image Captioning. The architecture improves both the image encoding and the language generation steps: it learns a multi-level representation of the relationships between image regions integrating learned a priori knowledge, and uses a mesh-like connectivity at decoding stage to exploit low- and high-level features. Experimentally, we investigate the performance of the M2 Transformer and different fully-attentive models in comparison with recurrent ones. When tested on COCO, our proposal achieves a new state of the art in single-model and ensemble configurations on the "Karpathy" test split and on the online test server. We also assess its performances when describing objects unseen in the training set. Trained models and code for reproducing the experiments are publicly available at: https://github.com/aimagelab/meshed-memory-transformer.
****

HCNAF: Hyper-Conditioned Neural Autoregressive Flow and its Application for Probabilistic Occupancy Map Forecasting
Geunseob Oh, Jean-Sebastien Valois; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14550-14559
We introduce Hyper-Conditioned Neural Autoregressive Flow (HCNAF); a powerful universal distribution approximator designed to model arbitrarily complex conditional probability density functions. HCNAF consists of a neural-net based conditional autoregressive flow (AF) and a hyper-network that can take large conditions

in non-autoregressive fashion and outputs the network parameters of the AF. Like other flow models, HCNAF performs exact likelihood inference. We conduct a number of density estimation tasks on toy experiments and MNIST to demonstrate the effectiveness and attributes of HCNAF, including its generalization capability over unseen conditions and expressivity. Finally, we show that HCNAF scales up to complex high-dimensional prediction problems of the magnitude of self-driving and that HCNAF yields a state-of-the-art performance in a public self-driving data set.
****

Non-Local Neural Networks With Grouped Bilinear Attentional Transforms
Lu Chi, Zehuan Yuan, Yadong Mu, Changhu Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11804-11813
Modeling spatial or temporal long-range dependency plays a key role in deep neural networks. Conventional dominant solutions include recurrent operations on sequential data or deeply stacking convolutional layers with small kernel size. Recently, a number of non-local operators (such as self-attention based) have been devised. They are typically generic and can be plugged into many existing network pipelines for globally computing among any two neurons in a feature map. This work proposes a novel non-local operator. It is inspired by the attention mechanism of human visual system, which can quickly attend to important local parts in sight and suppress other less-relevant information. The core of our method is learnable and data-adaptive bilinear attentional transform (BA-Transform), whose merits are three-folds: first, BA-Transform is versatile to model a wide spectrum of local or global attentional operations, such as emphasizing specific local regions. Each BA-Transform is learned in a data-adaptive way; Secondly, to address the discrepancy among features, we further design grouped BA-Transforms, which essentially apply different attentional operations to different groups of feature channels; Thirdly, many existing non-local operators are computation-intensive. The proposed BA-Transform is implemented by simple matrix multiplication and admits better efficacy. For empirical evaluation, we perform comprehensive experiments on two large-scale benchmarks, ImageNet and Kinetics, for image / video classification respectively. The achieved accuracies and various ablation experiments consistently demonstrate significant improvement by large margins.
****

Data-Free Knowledge Amalgamation via Group-Stack Dual-GAN
Jingwen Ye, Yixin Ji, Xinchao Wang, Xin Gao, Mingli Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12516-12525
Recent advances in deep learning have provided procedures for learning one network to amalgamate multiple streams of knowledge from the pre-trained Convolutional Neural Network (CNN) models, thus reduce the annotation cost. However, almost all existing methods demand massive training data, which may be unavailable due to privacy or transmission issues. In this paper, we propose a data-free knowledge amalgamate strategy to craft a well-behaved multi-task student network from multiple single/multi-task teachers. The main idea is to construct the group-stack generative adversarial networks (GANs) which have two dual generators. First one generator is trained to collect the knowledge by reconstructing the images approximating the original dataset utilized for pre-training the teachers. Then a dual generator is trained by taking the output from the former generator as input. Finally we treat the dual part generator as the target network and regroup it. As demonstrated on several benchmarks of multi-label classification, the proposed method without any training data achieves the surprisingly competitive results, even compared with some full-supervised methods.
****

JA-POLS: A Moving-Camera Background Model via Joint Alignment and Partially-Overlapping Local Subspaces
Irit Chelly, Vlad Winter, Dor Litvak, David Rosen, Oren Freifeld; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12585-12594
Background models are widely used in computer vision. While successful Static-ca

mera Background (SCB) models exist, Moving-camera Background (MCB) models are li
mited. Seemingly, there is a straightforward solution: 1) align the video frames
; 2) learn an SCB model; 3) warp either original or previously-unseen frames tow
ard the model. This approach, however, has drawbacks, especially when the accumu
lative camera motion is large and/or the video is long. Here we propose a purely
-2D unsupervised modular method that systematically eliminates those issues. Fir
st, to estimate warps in the original video, we solve a joint-alignment problem
while leveraging a certifiably-correct initialization. Next, we learn both multi
ple partially-overlapping local subspaces and how to predict alignments. Lastly,
 in test time, we warp a previously-unseen frame, based on the prediction, and p
roject it on a subset of those subspaces to obtain a background/foreground separ
ation. We show the method handles even large scenes with a relatively-free camer
a motion (provided the camera-to-scene distance does not change much) and that i
t not only yields State-of-the-Art results on the original video but also genera
lizes gracefully to previously-unseen videos of the same scene. Our code is avai
lable at https://github.com/BGU-CS-VIL/JA-POLS.
****

Mnemonics Training: Multi-Class Incremental Learning Without Forgetting
Yaoyao Liu, Yuting Su, An-An Liu, Bernt Schiele, Qianru Sun; Proceedings of
the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020,
 pp. 12245-12254
Multi-Class Incremental Learning (MCIL) aims to learn new concepts by incrementa
lly updating a model trained on previous concepts. However, there is an inherent
 trade-off to effectively learning new concepts without catastrophic forgetting
of previous ones. To alleviate this issue, it has been proposed to keep around a
 few examples of the previous concepts but the effectiveness of this approach he
avily depends on the representativeness of these examples. This paper proposes a
 novel and automatic framework we call mnemonics, where we parameterize exemplar
s and make them optimizable in an end-to-end manner. We train the framework thro
ugh bilevel optimizations, i.e., model-level and exemplar-level. We conduct exte
nsive experiments on three MCIL benchmarks, CIFAR-100, ImageNet-Subset and Image
Net, and show that using mnemonics exemplars can surpass the state-of-the-art by
 a large margin. Interestingly and quite intriguingly, the mnemonics exemplars t
end to be on the boundaries between different classes.
****

Orderless Recurrent Models for Multi-Label Classification
Vacit Oguz Yazici, Abel Gonzalez-Garcia, Arnau Ramisa, Bartlomiej Twardowski,
  Joost van de Weijer; Proceedings of the IEEE/CVF Conference on Computer Vision
 and Pattern Recognition (CVPR), 2020, pp. 13440-13449
Recurrent neural networks (RNN) are popular for many computer vision tasks, incl
uding multi-label classification. Since RNNs produce sequential outputs, labels
need to be ordered for the multi-label classification task. Current approaches s
ort labels according to their frequency, typically ordering them in either rare-
first or frequent-first. These imposed orderings do not take into account that t
he natural order to generate the labels can change for each image, e.g. first th
e dominant object before summing up the smaller objects in the image. Therefore,
 in this paper, we propose ways to dynamically order the ground truth labels wit
h the predicted label sequence. This allows for the faster training of more opti
mal LSTM models for multi-label classification. Analysis evidences that our meth
od does not suffer from duplicate generation, something which is common for othe
r models. Furthermore, it outperforms other CNN-RNN models, and we show that a s
tandard architecture of an image encoder and language decoder trained with our p
roposed loss obtains the state-of-the-art results on the challenging MS-COCO, WI
DER Attribute and PA-100K and competitive results on NUS-WIDE.
****

Exploring Category-Agnostic Clusters for Open-Set Domain Adaptation
Yingwei Pan, Ting Yao, Yehao Li, Chong-Wah Ngo, Tao Mei; Proceedings of the
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp.
 13867-13875
Unsupervised domain adaptation has received significant attention in recent year

s. Most of existing works tackle the closed-set scenario, assuming that the sour ce and target domains share the exactly same categories. In practice, neverthele ss, a target domain often contains samples of classes unseen in source domain (i .e., unknown class). The extension of domain adaptation from closed-set to such open-set situation is not trivial since the target samples in unknown class are not expected to align with the source. In this paper, we address this problem by augmenting the state-of-the-art domain adaptation technique, Self-Ensembling, w ith category-agnostic clusters in target domain. Specifically, we present Self-E nsembling with Category-agnostic Clusters (SE-CC) --- a novel architecture that steers domain adaptation with the additional guidance of category-agnostic clust ers that are specific to target domain. These clustering information provides do main-specific visual cues, facilitating the generalization of Self-Ensembling fo r both closed-set and open-set scenarios. Technically, clustering is firstly per formed over all the unlabeled target samples to obtain the category-agnostic clu sters, which reveal the underlying data space structure peculiar to target domai n. A clustering branch is capitalized on to ensure that the learnt representatio n preserves such underlying structure by matching the estimated assignment distr ibution over clusters to the inherent cluster distribution for each target sampl e. Furthermore, SE-CC enhances the learnt representation with mutual information maximization. Extensive experiments are conducted on Office and VisDA datasets for both open-set and closed-set domain adaptation, and superior results are rep orted when comparing to the state-of-the-art approaches.
****

Learning When and Where to Zoom With Deep Reinforcement Learning
Burak Uzkent, Stefano Ermon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12345-12354
While high resolution images contain semantically more useful information than t heir lower resolution counterparts, processing them is computationally more expe nsive, and in some applications, e.g. remote sensing, they can be much more expe nsive to acquire. For these reasons, it is desirable to develop an automatic met hod to selectively use high resolution data when necessary while maintaining acc uracy and reducing acquisition/run-time cost. In this direction, we propose Patc hDrop a reinforcement learning approach to dynamically identify when and where t o use/acquire high resolution data conditioned on the paired, cheap, low resolut ion images. We conduct experiments on CIFAR10, CIFAR100, ImageNet and fMoW datas ets where we use significantly less high resolution data while maintaining simil ar accuracy to models which use full high resolution images.
****

Densely Connected Search Space for More Flexible Neural Architecture Search
Jiemin Fang, Yuzhu Sun, Qian Zhang, Yuan Li, Wenyu Liu, Xinggang Wang; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2020, pp. 10628-10637
Neural architecture search (NAS) has dramatically advanced the development of ne ural network design. We revisit the search space design in most previous NAS met hods and find the number and widths of blocks are set manually. However, block c ounts and block widths determine the network scale (depth and width) and make a great influence on both the accuracy and the model cost (FLOPs/latency). In this paper, we propose to search block counts and block widths by designing a densel y connected search space, i.e., DenseNAS. The new search space is represented as a dense super network, which is built upon our designed routing blocks. In the super network, routing blocks are densely connected and we search for the best p ath between them to derive the final architecture. We further propose a chained cost estimation algorithm to approximate the model cost during the search. Both the accuracy and model cost are optimized in DenseNAS. For experiments on the Mo bileNetV2-based search space, DenseNAS achieves 75.3% top-1 accuracy on ImageNet with only 361MB FLOPs and 17.9ms latency on a single TITAN-XP. The larger model searched by DenseNAS achieves 76.1% accuracy with only 479M FLOPs. DenseNAS fur ther promotes the ImageNet classification accuracies of ResNet-18, -34 and -50-B by 1.5%, 0.5% and 0.3% with 200M, 600M and 680M FLOPs reduction respectively. T he related code is available at https://github.com/JaminFong/DenseNAS.

****
Neural Topological SLAM for Visual Navigation
Devendra Singh Chaplot, Ruslan Salakhutdinov, Abhinav Gupta, Saurabh Gupta; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12875-12884
This paper studies the problem of image-goal navigation which involves navigating to the location indicated by a goal image in a novel previously unseen environment. To tackle this problem, we design topological representations for space that effectively leverage semantics and afford approximate geometric reasoning. At the heart of our representations are nodes with associated semantic features, that are interconnected using coarse geometric information. We describe supervised learning-based algorithms that can build, maintain and use such representations under noisy actuation. Experimental study in visually and physically realistic simulation suggests that our method builds effective representations that capture structural regularities and efficiently solve long-horizon navigation problems. We observe a relative improvement of more than 50% over existing methods that study this task.
****
Generalized ODIN: Detecting Out-of-Distribution Image Without Learning From Out-of-Distribution Data
Yen-Chang Hsu, Yilin Shen, Hongxia Jin, Zsolt Kira; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10951-10960
Deep neural networks have attained remarkable performance when applied to data that comes from the same distribution as that of the training set, but can significantly degrade otherwise. Therefore, detecting whether an example is out-of-distribution (OoD) is crucial to enable a system that can reject such samples or alert users. Recent works have made significant progress on OoD benchmarks consisting of small image datasets. However, many recent methods based on neural networks rely on training or tuning with both in-distribution and out-of-distribution data. The latter is generally hard to define a-priori, and its selection can easily bias the learning. We base our work on a popular method ODIN, proposing two strategies for freeing it from the needs of tuning with OoD data, while improving its OoD detection performance. We specifically propose to decompose confidence scoring as well as a modified input pre-processing method. We show that both of these significantly help in detection performance. Our further analysis on a larger scale image dataset shows that the two types of distribution shifts, specifically semantic shift and non-semantic shift, present a significant difference in the difficulty of the problem, providing an analysis of when ODIN-like strategies do or do not work.
****
Sequential Motif Profiles and Topological Plots for Offline Signature Verification
Elias N. Zois, Evangelos Zervas, Dimitrios Tsourounis, George Economou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13248-13258
In spite of the overwhelming high-tech marvels and applications that rule our digital lives, the use of the handwritten signature is still recognized worldwide in government, personal and legal entities to be the most important behavioral biometric trait. A number of notable research approaches provide advanced results up to a certain point which allow us to assert with confidence that the performance attained by signature verification (SV) systems is comparable to those provided by any other biometric modality. Up to now, the mainstream trend for offline SV is shared between standard -or handcrafted- feature extraction methods and popular machine learning techniques, with typical examples ranging from sparse representation to Deep Learning. Recent progress in graph mining algorithms provide us with the prospect to re-evaluate the opportunity of utilizing graph representations by exploring corresponding graph features for offline SV. In this paper, inspired by the recent use of image visibility graphs for mapping images into networks, we introduce for the first time in offline SV literature their use as

a parameter free, agnostic representation for exploring global as well as local information. Global properties of the sparsely located content of the shape of the signature image are encoded with topological information of the whole graph. In addition, local pixel patches are encoded by sequential visibility motifs-subgraphs of size four, to a low six dimensional motif profile vector. A number of pooling functions operate on the motif codes in a spatial pyramid context in order to create the final feature vector. The effectiveness of the proposed method is evaluated with the use of two popular datasets. The local visibility graph features are considered to be highly informative for SV; this is sustained by the corresponding results which are at least comparable with other classic state-of-the-art approaches.

****

DOPS: Learning to Detect 3D Objects and Predict Their 3D Shapes

Mahyar Najibi, Guangda Lai, Abhijit Kundu, Zhichao Lu, Vivek Rathod, Thomas Funkhouser, Caroline Pantofaru, David Ross, Larry S. Davis, Alireza Fathi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11913-11922

We propose DOPS, a fast single-stage 3D object detection method for LIDAR data. Previous methods often make domain-specific design decisions, for example projecting points into a bird-eye view image in autonomous driving scenarios. In contrast, we propose a general-purpose method that works on both indoor and outdoor scenes. The core novelty of our method is a fast, single-pass architecture that both detects objects in 3D and estimates their shapes. 3D bounding box parameters are estimated in one pass for every point, aggregated through graph convolutions, and fed into a branch of the network that predicts latent codes representing the shape of each detected object. The latent shape space and shape decoder are learned on a synthetic dataset and then used as supervision for the end-to-end training of the 3D object detection pipeline. Thus our model is able to extract shapes without access to ground-truth shape information in the target dataset. During experiments, we find that our proposed method achieves state-of-the-art results by 5% on object detection in ScanNet scenes, and it gets top results by 3.4% in the Waymo Open Dataset, while reproducing the shapes of detected cars.

****

Multimodal Categorization of Crisis Events in Social Media

Mahdi Abavisani, Liwei Wu, Shengli Hu, Joel Tetreault, Alejandro Jaimes; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14679-14689

Recent developments in image classification and natural language processing, coupled with the rapid growth in social media usage, have enabled fundamental advances in detecting breaking events around the world in real-time. Emergency response is one such area that stands to gain from these advances. By processing billions of texts and images a minute, events can be automatically detected to enable emergency response workers to better assess rapidly evolving situations and deploy resources accordingly. To date, most event detection techniques in this area have focused on image-only or text-only approaches, limiting detection performance and impacting the quality of information delivered to crisis response teams. In this paper, we present a new multimodal fusion method that leverages both images and texts as input. In particular, we introduce a cross-attention module that can filter uninformative and misleading components from weak modalities on a sample by sample basis. In addition, we employ a multimodal graph-based approach to stochastically transition between embeddings of different multimodal pairs during training to better regularize the learning process as well as dealing with limited training data by constructing new matched pairs from different samples. We show that our method outperforms the unimodal approaches and strong multimodal baselines by a large margin on three crisis-related tasks.

****

Physically Realizable Adversarial Examples for LiDAR Object Detection

James Tu, Mengye Ren, Sivabalan Manivasagam, Ming Liang, Bin Yang, Richard Du, Frank Cheng, Raquel Urtasun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13716-13725

Modern autonomous driving systems rely heavily on deep learning models to proces s point cloud sensory data; meanwhile, deep models have been shown to be suscept ible to adversarial attacks with visually imperceptible perturbations. Despite t he fact that this poses a security concern for the self-driving industry, there has been very little exploration in terms of 3D perception, as most adversarial attacks have only been applied to 2D flat images. In this paper, we address this issue and present a method to generate universal 3D adversarial objects to fool LiDAR detectors. In particular, we demonstrate that placing an adversarial obje ct on the rooftop of any target vehicle to hide the vehicle entirely from LiDAR detectors with a success rate of 80%. We report attack results on a suite of det ectors using various input representation of point clouds. We also conduct a pil ot study on adversarial defense using data augmentation. This is one step closer towards safer self-driving under unseen conditions from limited training data.
****

STEFANN: Scene Text Editor Using Font Adaptive Neural Network
Prasun Roy, Saumik Bhattacharya, Subhankar Ghosh, Umapada Pal; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020 , pp. 13228-13237
Textual information in a captured scene plays an important role in scene interpr etation and decision making. Though there exist methods that can successfully de tect and interpret complex text regions present in a scene, to the best of our k nowledge, there is no significant prior work that aims to modify the textual inf ormation in an image. The ability to edit text directly on images has several ad vantages including error correction, text restoration and image reusability. In this paper, we propose a method to modify text in an image at character-level. W e approach the problem in two stages. At first, the unobserved character (target) is generated from an observed character (source) being modified. We propose tw o different neural network architectures - (a) FANnet to achieve structural cons istency with source font and (b) Colornet to preserve source color. Next, we rep lace the source character with the generated character maintaining both geometri c and visual consistency with neighboring characters. Our method works as a unif ied platform for modifying text in images. We present the effectiveness of our m ethod on COCO-Text and ICDAR datasets both qualitatively and quantitatively.
****

SwapText: Image Based Texts Transfer in Scenes
Qiangpeng Yang, Jun Huang, Wei Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14700-14709
Swapping text in scene images while preserving original fonts, colors, sizes and background textures is a challenging task due to the complex interplay between different factors. In this work, we present SwapText, a three-stage framework to transfer texts across scene images. First, a novel text swapping network is pro posed to replace text labels only in the foreground image. Second, a background completion network is learned to reconstruct background images. Finally, the gen erated foreground image and background image are used to generate the word image by the fusion network. Using the proposing framework, we can manipulate the tex ts of the input images even with severe geometric distortion. Qualitative and qu antitative results are presented on several scene text datasets, including regul ar and irregular text datasets. We conducted extensive experiments to prove the usefulness of our method such as image based text translation, text image synthe sis.
****

MetaFuse: A Pre-trained Fusion Model for Human Pose Estimation
Rongchang Xie, Chunyu Wang, Yizhou Wang; Proceedings of the IEEE/CVF Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13686-13695
Cross view feature fusion is the key to address the occlusion problem in human p ose estimation. The current fusion methods need to train a separate model for ev ery pair of cameras making them difficult to scale. In this work, we introduce M etaFuse, a pre-trained fusion model learned from a large number of cameras in th e Panoptic dataset. The model can be efficiently adapted or finetuned for a new pair of cameras using a small number of labeled images. The strong adaptation po

wer of MetaFuse is due in large part to the proposed factorization of the original fusion model into two parts--(1) a generic fusion model shared by all cameras, and (2) lightweight camera-dependent transformations. Furthermore, the generic model is learned from many cameras by a meta-learning style algorithm to maximize its adaptation capability to various camera poses. We observe in experiments that MetaFuse finetuned on the public datasets outperforms the state-of-the-arts by a large margin which validates its value in practice.
****

Local-Global Video-Text Interactions for Temporal Grounding
Jonghwan Mun, Minsu Cho, Bohyung Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10810-10819
This paper addresses the problem of text-to-video temporal grounding, which aims to identify the time interval in a video semantically relevant to a text query. We tackle this problem using a novel regression-based model that learns to extract a collection of mid-level features for semantic phrases in a text query, which corresponds to important semantic entities described in the query (e.g., actors, objects, and actions), and reflect bi-modal interactions between the linguistic features of the query and the visual features of the video in multiple levels. The proposed method effectively predicts the target time interval by exploiting contextual information from local to global during bi-modal interactions. Through in-depth ablation studies, we find out that incorporating both local and global context in video and text interactions is crucial to the accurate grounding. Our experiment shows that the proposed method outperforms the state of the arts on Charades-STA and ActivityNet Captions datasets by large margins, 7.44% and 4.61% points at Recall@tIoU=0.5 metric, respectively.
****

MotionNet: Joint Perception and Motion Prediction for Autonomous Driving Based on Bird's Eye View Maps
Pengxiang Wu, Siheng Chen, Dimitris N. Metaxas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11385-11395
The ability to reliably perceive the environmental states, particularly the existence of objects and their motion behavior, is crucial for autonomous driving. In this work, we propose an efficient deep model, called MotionNet, to jointly perform perception and motion prediction from 3D point clouds. MotionNet takes a sequence of LiDAR sweeps as input and outputs a bird's eye view (BEV) map, which encodes the object category and motion information in each grid cell. The backbone of MotionNet is a novel spatio-temporal pyramid network, which extracts deep spatial and temporal features in a hierarchical fashion. To enforce the smoothness of predictions over both space and time, the training of MotionNet is further regularized with novel spatial and temporal consistency losses. Extensive experiments show that the proposed method overall outperforms the state-of-the-arts, including the latest scene-flow- and 3D-object-detection-based methods. This indicates the potential value of the proposed method serving as a backup to the bounding-box-based system, and providing complementary information to the motion planner in autonomous driving. Code is available at https://www.merl.com/research/license#MotionNet.
****

Improving Action Segmentation via Graph-Based Temporal Reasoning
Yifei Huang, Yusuke Sugano, Yoichi Sato; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14024-14034
Temporal relations among multiple action segments play an important role in action segmentation especially when observations are limited (e.g., actions are occluded by other objects or happen outside a field of view). In this paper, we propose a network module called Graph-based Temporal Reasoning Module (GTRM) that can be built on top of existing action segmentation models to learn the relation of multiple action segments in various time spans. We model the relations by using two Graph Convolution Networks (GCNs) where each node represents an action segment. The two graphs have different edge properties to account for boundary regression and classification tasks, respectively. By applying graph convolution, we

can update each node's representation based on its relation with neighboring nodes. The updated representation is then used for improved action segmentation. We evaluate our model on the challenging egocentric datasets namely EGTEA and EPIC-Kitchens, where actions may be partially observed due to the viewpoint restriction. The results show that our proposed GTRM outperforms state-of-the-art action segmentation models by a large margin. We also demonstrate the effectiveness of our model on two third-person video datasets, the 50Salads dataset and the Breakfast dataset.

****

DeepEMD: Few-Shot Image Classification With Differentiable Earth Mover's Distance and Structured Classifiers

Chi Zhang, Yujun Cai, Guosheng Lin, Chunhua Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12203-12213

In this paper, we address the few-shot classification task from a new perspective of optimal matching between image regions. We adopt the Earth Mover's Distance (EMD) as a metric to compute a structural distance between dense image representations to determine image relevance. The EMD generates the optimal matching flows between structural elements that have the minimum matching cost, which is used to represent the image distance for classification. To generate the important weights of elements in the EMD formulation, we design a cross-reference mechanism, which can effectively minimize the impact caused by the cluttered background and large intra-class appearance variations. To handle k-shot classification, we propose to learn a structured fully connected layer that can directly classify dense image representations with the EMD. Based on the implicit function theorem, the EMD can be inserted as a layer into the network for end-to-end training. We conduct comprehensive experiments to validate our algorithm and we set new state-of-the-art performance on four popular few-shot classification benchmarks, namely miniImageNet, tieredImageNet, Fewshot-CIFAR100 (FC100) and Caltech-UCSD Birds-200-2011 (CUB).

****

Conditional Gaussian Distribution Learning for Open Set Recognition

Xin Sun, Zhenning Yang, Chi Zhang, Keck-Voon Ling, Guohao Peng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13480-13489

Deep neural networks have achieved state-of-the-art performance in a wide range of recognition/classification tasks. However, when applying deep learning to real-world applications, there are still multiple challenges. A typical challenge is that unknown samples may be fed into the system during the testing phase and traditional deep neural networks will wrongly recognize the unknown sample as one of the known classes. Open set recognition is a potential solution to overcome this problem, where the open set classifier should have the ability to reject unknown samples as well as maintain high classification accuracy on known classes. The variational auto-encoder (VAE) is a popular model to detect unknowns, but it cannot provide discriminative representations for known classification. In this paper, we propose a novel method, Conditional Gaussian Distribution Learning (CGDL), for open set recognition. In addition to detecting unknown samples, this method can also classify known samples by forcing different latent features to approximate different Gaussian models. Meanwhile, to avoid information hidden in the input vanishing in the middle layers, we also adopt the probabilistic ladder architecture to extract high-level abstract features. Experiments on several standard image datasets reveal that the proposed method significantly outperforms the baseline method and achieves new state-of-the-art results.

****

D2Det: Towards High Quality Object Detection and Instance Segmentation

Jiale Cao, Hisham Cholakkal, Rao Muhammad Anwer, Fahad Shahbaz Khan, Yanwei Pang, Ling Shao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11485-11494

We propose a novel two-stage detection method, D2Det, that collectively addresses both precise localization and accurate classification. For precise localizatio

n, we introduce a dense local regression that predicts multiple dense box offset
s for an object proposal. Different from traditional regression and keypoint-bas
ed localization employed in two-stage detectors, our dense local regression is n
ot limited to a quantized set of keypoints within a fixed region and has the abi
lity to regress position-sensitive real number dense offsets, leading to more pr
ecise localization. The dense local regression is further improved by a binary o
verlap prediction strategy that reduces the influence of background region on th
e final box regression. For accurate classification, we introduce a discriminati
ve RoI pooling scheme that samples from various sub-regions of a proposal and pe
rforms adaptive weighting to obtain discriminative features. On MS COCO test-dev
, our D2Det outperforms existing two-stage methods, with a single-model performa
nce of 45.4 AP, using ResNet101 backbone. When using multi-scale training and in
ference, D2Det obtains AP of 50.1. In addition to detection, we adapt D2Det for
instance segmentation, achieving a mask AP of 40.2 with a two-fold speedup, comp
ared to the state-of-the-art. We also demonstrate the effectiveness of our D2Det
on airborne sensors by performing experiments for object detection in UAV image
s (UAVDT dataset) and instance segmentation in satellite images (iSAID dataset).
Source code is available at https://github.com/JialeCao001/D2Det.
****

Rethinking Classification and Localization for Object Detection
Yue Wu, Yinpeng Chen, Lu Yuan, Zicheng Liu, Lijuan Wang, Hongzhi Li, Yun F
u; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni
tion (CVPR), 2020, pp. 10186-10195
Two head structures (i.e. fully connected head and convolution head) have been w
idely used in R-CNN based detectors for classification and localization tasks. H
owever, there is a lack of understanding of how does these two head structures w
ork for these two tasks. To address this issue, we perform a thorough analysis a
nd find an interesting fact that the two head structures have opposite preferenc
es towards the two tasks. Specifically, the fully connected head (fc-head) is mo
re suitable for the classification task, while the convolution head (conv-head)
is more suitable for the localization task. Furthermore, we examine the output f
eature maps of both heads and find that fc-head has more spatial sensitivity tha
n conv-head. Thus, fc-head has more capability to distinguish a complete object
from part of an object, but is not robust to regress the whole object. Based upo
n these findings, we propose a Double-Head method, which has a fully connected h
ead focusing on classification and a convolution head for bounding box regressio
n. Without bells and whistles, our method gains +3.5 and +2.8 AP on MS COCO data
set from Feature Pyramid Network (FPN) baselines with ResNet-50 and ResNet-101 b
ackbones, respectively.
****

Two-Stage Peer-Regularized Feature Recombination for Arbitrary Image Style Trans
fer
Jan Svoboda, Asha Anoosheh, Christian Osendorfer, Jonathan Masci; Proceedings
of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2
020, pp. 13816-13825
This paper introduces a neural style transfer model to generate a stylized image
conditioning on a set of examples describing the desired style. The proposed so
lution produces high-quality images even in the zero-shot setting and allows for
more freedom in changes to the content geometry. This is made possible by intro
ducing a novel Two-Stage Peer-Regularization Layer that recombines style and con
tent in latent space by means of a custom graph convolutional layer. Contrary to
the vast majority of existing solutions, our model does not depend on any pre-t
rained networks for computing perceptual losses and can be trained fully end-to-
end thanks to a new set of cyclic losses that operate directly in latent space a
nd not on the RGB images. An extensive ablation study confirms the usefulness of
the proposed losses and of the Two-Stage Peer-Regularization Layer, with qualit
ative results that are competitive with respect to the current state of the art
using a single model for all presented styles. This opens the door to more abstr
act and artistic neural image generation scenarios, along with simpler deploymen
t of the model.

****
Synthetic Learning: Learn From Distributed Asynchronized Discriminator GAN Without Sharing Medical Image Data
Qi Chang, Hui Qu, Yikai Zhang, Mert Sabuncu, Chao Chen, Tong Zhang, Dimitris N. Metaxas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13856-13866
In this paper, we propose a data privacy-preserving and communication efficient distributed GAN learning framework named Distributed Asynchronized Discriminator GAN (AsynDGAN). Our proposed framework aims to train a central generator learns from distributed discriminator, and use the generated synthetic image solely to train the segmentation model. We validate the proposed framework on the application of health entities learning problem which is known to be privacy sensitive. Our experiments show that our approach: 1) could learn the real image's distribution from multiple datasets without sharing the patient's raw data. 2) is more efficient and requires lower bandwidth than other distributed deep learning methods. 3) achieves higher performance compared to the model trained by one real dataset, and almost the same performance compared to the model trained by all real datasets. 4) has provable guarantees that the generator could learn the distributed distribution in an all important fashion thus is unbiased.We release our AsynDGAN source code at: https://github.com/tommy-qichang/AsynDGAN
****
BEDSR-Net: A Deep Shadow Removal Network From a Single Document Image
Yun-Hsuan Lin, Wen-Chin Chen, Yung-Yu Chuang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12905-12914
Removing shadows in document images enhances both the visual quality and readability of digital copies of documents. Most existing shadow removal algorithms for document images use hand-crafted heuristics and are often not robust to documents with different characteristics. This paper proposes the Background Estimation Document Shadow Removal Network (BEDSR-Net), the first deep network specifically designed for document image shadow removal. For taking advantage of specific properties of document images, a background estimation module is designed for extracting the global background color of the document. During the process of estimating the background color, the module also learns information about the spatial distribution of background and non-background pixels. We encode such information into an attention map. With the estimated global background color and attention map, the shadow removal network can better recover the shadow-free image. We also show that the model trained on synthetic images remains effective for real photos, and provide a large set of synthetic shadow images of documents along with their corresponding shadow-free images and shadow masks. Extensive quantitative and qualitative experiments on several benchmarks show that the BEDSR-Net outperforms existing methods in enhancing both the visual quality and readability of document images.
****
Label Decoupling Framework for Salient Object Detection
Jun Wei, Shuhui Wang, Zhe Wu, Chi Su, Qingming Huang, Qi Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13025-13034
To get more accurate saliency maps, recent methods mainly focus on aggregating multi-level features from fully convolutional network (FCN) and introducing edge information as auxiliary supervision. Though remarkable progress has been achieved, we observe that the closer the pixel is to the edge, the more difficult it is to be predicted, because edge pixels have a very imbalance distribution. To address this problem, we propose a label decoupling framework (LDF) which consists of a label decoupling (LD) procedure and a feature interaction network (FIN). LD explicitly decomposes the original saliency map into body map and detail map, where body map concentrates on center areas of objects and detail map focuses on regions around edges. Detail map works better because it involves much more pixels than traditional edge supervision. Different from saliency map, body map discards edge pixels and only pays attention to center areas. This successfully avoids the distraction from edge pixels during training. Therefore, we employ two b

ranches in FIN to deal with body map and detail map respectively. Feature interaction (FI) is designed to fuse the two complementary branches to predict the saliency map, which is then used to refine the two branches again. This iterative refinement is helpful for learning better representations and more precise saliency maps. Comprehensive experiments on six benchmark datasets demonstrate that LDF outperforms state-of-the-art approaches on different evaluation metrics.
****

LG-GAN: Label Guided Adversarial Network for Flexible Targeted Attack of Point Cloud Based Deep Networks
Hang Zhou, Dongdong Chen, Jing Liao, Kejiang Chen, Xiaoyi Dong, Kunlin Liu, Weiming Zhang, Gang Hua, Nenghai Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10356-10365
Deep neural networks have made tremendous progress in 3D point-cloud recognition. Recent works have shown that these 3D recognition networks are also vulnerable to adversarial samples produced from various attack methods, including optimization-based 3D Carlini-Wagner attack, gradient-based iterative fast gradient method, and skeleton-detach based point-dropping. However, after a careful analysis, these methods are either extremely slow because of the optimization/iterative scheme, or not flexible to support targeted attack of a specific category. To overcome these shortcomings, this paper proposes a novel label guided adversarial network (LG-GAN) for real-time flexible targeted point cloud attack. To the best of our knowledge, this is the first generation based 3D point cloud attack method. By feeding the original point clouds and target attack label into LG-GAN, it can learn how to deform the point clouds to mislead the recognition network into the specific label only with a single forward pass. In detail, LG-GAN first leverages one multi-branch adversarial network to extract hierarchical features of the input point clouds, then incorporates the specified label information into multiple intermediate features using the label encoder. Finally, the encoded features will be fed into the coordinate reconstruction decoder to generate the target adversarial sample. By evaluating different point-cloud recognition models (e.g., PointNet, PointNet++ and DGCNN), we demonstrate that the proposed LG-GAN can support flexible targeted attack on the fly while guaranteeing good attack performance and higher efficiency simultaneously.
****

Look-Into-Object: Self-Supervised Structure Modeling for Object Recognition
Mohan Zhou, Yalong Bai, Wei Zhang, Tiejun Zhao, Tao Mei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11774-11783
Most object recognition approaches predominantly focus on learning discriminative visual patterns, while overlooking the holistic object structure. Though important, structure modeling usually requires significant manual annotations and therefore is labor-intensive. In this paper, we propose to "look into object" (explicitly yet intrinsically model the object structure) through incorporating self-supervisions into the traditional framework. We show the recognition backbone can be substantially enhanced for more robust representation learning, without any cost of extra annotation and inference speed. Specifically, we first propose an object-extent learning module for localizing the object according to the visual patterns shared among the instances in the same category. We then design a spatial context learning module for modeling the internal structures of the object, through predicting the relative positions within the extent. These two modules can be easily plugged into any backbone networks during training and detached at inference time. Extensive experiments show that our look-into-object approach (LIO) achieves large performance gain on a number of benchmarks, including generic object recognition (ImageNet) and fine-grained object recognition tasks (CUB, Cars, Aircraft). We also show that this learning paradigm is highly generalizable to other tasks such as object detection and segmentation (MS COCO). Project page: https://github.com/JDAI-CV/LIO.
****

Inferring Attention Shift Ranks of Objects for Image Saliency
Avishek Siris, Jianbo Jiao, Gary K.L. Tam, Xianghua Xie, Rynson W.H. Lau; Pr

oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12133-12143
Psychology studies and behavioural observation show that humans shift their attention from one location to another when viewing an image of a complex scene. This is due to the limited capacity of the human visual system in simultaneously processing multiple visual inputs. The sequential shifting of attention on objects in a non-task oriented viewing can be seen as a form of saliency ranking. Although there are methods proposed for predicting saliency rank, they are not able to model this human attention shift well, as they are primarily based on ranking saliency values from binary prediction. Following psychological studies, in this paper, we propose to predict the saliency rank by inferring human attention shift. Due to the lack of such data, we first construct a large-scale salient object ranking dataset. The saliency rank of objects is defined by the order that an observer attends to these objects based on attention shift. The final saliency rank is an average across the saliency ranks of multiple observers. We then propose a learning-based CNN to leverage both bottom-up and top-down attention mechanisms to predict the saliency rank. Experimental results show that the proposed network achieves state-of-the-art performances on salient object rank prediction. Code and dataset are available at https://github.com/SirisAvishek/Attention_Shift_Ranks
****
Music Gesture for Visual Sound Separation
Chuang Gan, Deng Huang, Hang Zhao, Joshua B. Tenenbaum, Antonio Torralba; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10478-10487
Recent deep learning approaches have achieved impressive performance on visual sound separation tasks. However, these approaches are mostly built on appearance and optical flow like motion feature representations, which exhibit limited abilities to find the correlations between audio signals and visual points, especially when separating multiple instruments of the same types, such as multiple violins in a scene. To address this, we propose "Music Gesture," a keypoint-based structured representation to explicitly model the body and finger movements of musicians when they perform music. We first adopt a context-aware graph network to integrate visual semantic context with body dynamics and then apply an audio-visual fusion model to associate body movements with the corresponding audio signals. Experimental results on three music performance datasets show: 1) strong improvements upon benchmark metrics for hetero-musical separation tasks (i.e. different instruments); 2) new ability for effective homo-musical separation for piano, flute, and trumpet duets, which to our best knowledge has never been achieved with alternative methods.
****
Exploring Categorical Regularization for Domain Adaptive Object Detection
Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, Xiu-Shen Wei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11724-11733
In this paper, we tackle the domain adaptive object detection problem, where the main challenge lies in significant domain gaps between source and target domains. Previous work seeks to plainly align image-level and instance-level shifts to eventually minimize the domain discrepancy. However, they still overlook to match crucial image regions and important instances across domains, which will strongly affect domain shift mitigation. In this work, we propose a simple but effective categorical regularization framework for alleviating this issue. It can be applied as a plug-and-play component on a series of Domain Adaptive Faster R-CNN methods which are prominent for dealing with domain adaptive detection. Specifically, by integrating an image-level multi-label classifier upon the detection backbone, we can obtain the sparse but crucial image regions corresponding to categorical information, thanks to the weakly localization ability of the classification manner. Meanwhile, at the instance level, we leverage the categorical consistency between image-level predictions (by the classifier) and instance-level predictions (by the detection head) as a regularization factor to automatically h

unt for the hard aligned instances of target domains. Extensive experiments of various domain shift scenarios show that our method obtains a significant performance gain over original Domain Adaptive Faster R-CNN detectors. Furthermore, qualitative visualization and analyses can demonstrate the ability of our method for attending on the key regions/instances targeting on domain adaptation. Our code is open-source and available at https://github.com/Megvii-Nanjing/CR-DA-DET.
****

Incremental Learning in Online Scenario
Jiangpeng He, Runyu Mao, Zeman Shao, Fengqing Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13926-13935
Modern deep learning approaches have achieved great success in many vision applications by training a model using all available task-specific data. However, there are two major obstacles making it challenging to implement for real life applications: (1) Learning new classes makes the trained model quickly forget old classes knowledge, which is referred to as catastrophic forgetting. (2) As new observations of old classes come sequentially over time, the distribution may change in unforeseen way, making the performance degrade dramatically on future data, which is referred to as concept drift. Current state-of-the-art incremental learning methods require a long time to train the model whenever new classes are added and none of them takes into consideration the new observations of old classes. In this paper, we propose an incremental learning framework that can work in the challenging online learning scenario and handle both new classes data and new observations of old classes. We address problem (1) in online mode by introducing a modified cross-distillation loss together with a two-step learning technique. Our method outperforms the results obtained from current state-of-the-art offline incremental learning methods on the CIFAR-100 and ImageNet-1000 (ILSVRC 2012) datasets under the same experiment protocol but in online scenario. We also provide a simple yet effective method to mitigate problem (2) by updating exemplar set using the feature of each new observation of old classes and demonstrate a real life application of online food image classification based on our complete framework using the Food-101 dataset.
****

Gait Recognition via Semi-supervised Disentangled Representation Learning to Identity and Covariate Features
Xiang Li, Yasushi Makihara, Chi Xu, Yasushi Yagi, Mingwu Ren; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13309-13319
Existing gait recognition approaches typically focus on learning identity features that are invariant to covariates (e.g., the carrying status, clothing, walking speed, and viewing angle) and seldom involve learning features from the covariate aspect, which may lead to failure modes when variations due to the covariate overwhelm those due to the identity. We therefore propose a method of gait recognition via disentangled representation learning that considers both identity and covariate features. Specifically, we first encode an input gait template to get the disentangled identity and covariate features, and then decode the features to simultaneously reconstruct the input gait template and the canonical version of the same subject with no covariates in a semi-supervised manner to ensure successful disentanglement. We finally feed the disentangled identity features into a contrastive/triplet loss function for a verification/identification task. Moreover, we find that new gait templates can be synthesized by transferring the covariate feature from one subject to another. Experimental results on three publicly available gait data sets demonstrate the effectiveness of the proposed method compared with other state-of-the-art methods.
****

Discovering Human Interactions With Novel Objects via Zero-Shot Learning
Suchen Wang, Kim-Hui Yap, Junsong Yuan, Yap-Peng Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11652-11661
We aim to detect human interactions with novel objects through zero-shot learnin

g. Different from previous works, we allow unseen object categories by using its semantic word embedding. To do so, we design a human-object region proposal network specifically for the human-object interaction detection task. The core idea is to leverage human visual clues to localize objects which are interacting with humans. We show that our proposed model can outperform existing methods on detecting interacting objects, and generalize well to novel objects. To recognize objects from unseen categories, we devise a zero-shot classification module upon the classifier of seen categories. It utilizes the classifier logits for seen categories to estimate a vector in the semantic space, and then performs nearest search to find the closest unseen category. We validate our method on V-COCO and HICO-DET datasets, and obtain superior results on detecting human interactions with both seen and unseen objects.
****

Visual Commonsense R-CNN
Tan Wang, Jianqiang Huang, Hanwang Zhang, Qianru Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10760-10770
We present a novel unsupervised feature representation learning method, Visual Commonsense Region-based Convolutional Neural Network (VC R-CNN), to serve as an improved visual region encoder for high-level tasks such as captioning and VQA. Given a set of detected object regions in an image (e.g., using Faster R-CNN), like any other unsupervised feature learning methods (e.g., word2vec), the proxy training objective of VC R-CNN is to predict the contextual objects of a region. However, they are fundamentally different: the prediction of VC R-CNN is by using causal intervention: P(Y|do(X)), while others are by using the conventional likelihood: P(Y|X). This is also the core reason why VC R-CNN can learn "sense-making" knowledge like chair can be sat -- while not just "common" co-occurrences such as chair is likely to exist if table is observed. We extensively apply VC R-CNN features in prevailing models of three popular tasks: Image Captioning, VQA, and VCR, and observe consistent performance boosts across them, achieving many new state-of-the-arts.
****

Squeeze-and-Attention Networks for Semantic Segmentation
Zilong Zhong, Zhong Qiu Lin, Rene Bidart, Xiaodan Hu, Ibrahim Ben Daya, Zhifeng Li, Wei-Shi Zheng, Jonathan Li, Alexander Wong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13065-13074
The recent integration of attention mechanisms into segmentation networks improves their representational capabilities through a great emphasis on more informative features. However, these attention mechanisms ignore an implicit sub-task of semantic segmentation and are constrained by the grid structure of convolution kernels. In this paper, we propose a novel squeeze-and-attention network (SANet) architecture that leverages an effective squeeze-and-attention (SA) module to account for two distinctive characteristics of segmentation: i) pixel-group attention, and ii) pixel-wise prediction. Specifically, the proposed SA modules impose pixel-group attention on conventional convolution by introducing an 'attention' convolutional channel, thus taking into account spatial-channel inter-dependencies in an efficient manner. The final segmentation results are produced by merging outputs from four hierarchical stages of a SANet to integrate multi-scale contexts for obtaining an enhanced pixel-wise prediction. Empirical experiments on two challenging public datasets validate the effectiveness of the proposed SANets, which achieves 83.2 % mIoU (without COCO pre-training) on PASCAL VOC and a state-of-the-art mIoU of 54.4 % on PASCAL Context.
****

UNAS: Differentiable Architecture Search Meets Reinforcement Learning
Arash Vahdat, Arun Mallya, Ming-Yu Liu, Jan Kautz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11266-11275
Neural architecture search (NAS) aims to discover network architectures with desired properties such as high accuracy or low latency. Recently, differentiable N

AS (DNAS) has demonstrated promising results while maintaining a search cost ord ers of magnitude lower than reinforcement learning (RL) based NAS. However, DNAS models can only optimize differentiable loss functions in search, and they requ ire an accurate differentiable approximation of non-differentiable criteria. In this work, we present UNAS, a unified framework for NAS, that encapsulates recen t DNAS and RL-based approaches under one framework. Our framework brings the bes t of both worlds, and it enables us to search for architectures with both differ entiable and non-differentiable criteria in one unified framework while maintain ing a low search cost. Further, we introduce a new objective function for search based on the generalization gap that prevents the selection of architectures pr one to overfitting. We present extensive experiments on the CIFAR-10, CIFAR-100 and ImageNet datasets and we perform search in two fundamentally different searc h spaces. We show that UNAS obtains the state-of-the-art average accuracy on all three datasets when compared to the architectures searched in the DARTS space. Moreover, we show that UNAS can find an efficient and accurate architecture in t he ProxylessNAS search space, that outperforms existing MobileNetV2 based archit ectures. The source code is available at https://github.com/NVlabs/unas.

****

Multiple Anchor Learning for Visual Object Detection

Wei Ke, Tianliang Zhang, Zeyi Huang, Qixiang Ye, Jianzhuang Liu, Dong Huang ; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit ion (CVPR), 2020, pp. 10206-10215

Classification and localization are two pillars of visual object detectors. Howe ver, in CNN-based detectors, these two modules are usually optimized under a fix ed set of candidate (or anchor) bounding boxes. This configuration significantly limits the possibility to jointly optimize classification and localization. In this paper, we propose a Multiple Instance Learning (MIL) approach that selects anchors and jointly optimizes the two modules of a CNN-based object detector. Ou r approach, referred to as Multiple Anchor Learning (MAL), constructs anchor bag s and selects the most representative anchors from each bag. Such an iterative s election process is potentially NP-hard to optimize. To address this issue, we s olve MAL by repetitively depressing the confidence of selected anchors by pertur bing their corresponding features. In an adversarial selection-depression manner , MAL not only pursues optimal solutions but also fully leverages multiple ancho rs/features to learn a detection model. Experiments show that MAL improves the b aseline RetinaNet with significant margins on the commonly used MS-COCO object d etection benchmark and achieves new state-of-the-art detection performance compa red with recent methods.

****

Explaining Knowledge Distillation by Quantifying the Knowledge

Xu Cheng, Zhefan Rao, Yilan Chen, Quanshi Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12925-12 935

This paper presents a method to interpret the success of knowledge distillation by quantifying and analyzing task-relevant and task-irrelevant visual concepts t hat are encoded in intermediate layers of a deep neural network (DNN). More spec ifically, three hypotheses are proposed as follows. 1. Knowledge distillation ma kes the DNN learn more visual concepts than learning from raw data. 2. Knowledge distillation ensures that the DNN is prone to learning various visual concepts simultaneously. Whereas, in the scenario of learning from raw data, the DNN lear ns visual concepts sequentially. 3. Knowledge distillation yields more stable op timization directions than learning from raw data. Accordingly, we design three types of mathematical metrics to evaluate feature representations of the DNN. In experiments, we diagnosed various DNNs, and above hypotheses were verified.

****

Unifying Training and Inference for Panoptic Segmentation

Qizhu Li, Xiaojuan Qi, Philip H.S. Torr; Proceedings of the IEEE/CVF Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13320-13328

We present an end-to-end network to bridge the gap between training and inferenc e pipeline for panoptic segmentation, a task that seeks to partition an image in

to semantic regions for "stuff" and object instances for "things". In contrast t
o recent works, our network exploits a parametrised, yet lightweight panoptic se
gmentation submodule, powered by an end-to-end learnt dense instance affinity, t
o capture the probability that any pair of pixels belong to the same instance. T
his panoptic submodule gives rise to a novel propagation mechanism for panoptic
logits and enables the network to output a coherent panoptic segmentation map fo
r both "stuff" and "thing" classes, without any post-processing. Reaping the ben
efits of end-to-end training, our full system sets new records on the popular st
reet scene dataset, Cityscapes, achieving 61.4 PQ with a ResNet-50 backbone usin
g only the fine annotations. On the challenging COCO dataset, our ResNet-50-base
d network also delivers state-of-the-art accuracy of 43.4 PQ. Moreover, our netw
ork flexibly works with and without object mask cues, performing competitively u
nder both settings, which is of interest for applications with computation budge
ts.
****

Scalable Uncertainty for Computer Vision With Functional Variational Inference
Eduardo D. C. Carvalho, Ronald Clark, Andrea Nicastro, Paul H. J. Kelly; Proc
eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C
VPR), 2020, pp. 12003-12013
As Deep Learning continues to yield successful applications in Computer Vision,
the ability to quantify all forms of uncertainty is a paramount requirement for
its safe and reliable deployment in the real-world. In this work, we leverage th
e formulation of variational inference in function space, where we associate Gau
ssian Processes (GPs) to both Bayesian CNN priors and variational family. Since
GPs are fully determined by their mean and covariance functions, we are able to
obtain predictive uncertainty estimates at the cost of a single forward pass thr
ough any chosen CNN architecture and for any supervised learning task. By levera
ging the structure of the induced covariance matrices, we propose numerically ef
ficient algorithms which enable fast training in the context of high-dimensional
tasks such as depth estimation and semantic segmentation. Additionally, we prov
ide sufficient conditions for constructing regression loss functions whose proba
bilistic counterparts are compatible with aleatoric uncertainty quantification.
****

X-Linear Attention Networks for Image Captioning
Yingwei Pan, Ting Yao, Yehao Li, Tao Mei; Proceedings of the IEEE/CVF Confere
nce on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10971-10980
Recent progress on fine-grained visual recognition and visual question answering
has featured Bilinear Pooling, which effectively models the 2nd order interacti
ons across multi-modal inputs. Nevertheless, there has not been evidence in supp
ort of building such interactions concurrently with attention mechanism for imag
e captioning. In this paper, we introduce a unified attention block --- X-Linear
attention block, that fully employs bilinear pooling to selectively capitalize
on visual information or perform multi-modal reasoning. Technically, X-Linear at
tention block simultaneously exploits both the spatial and channel-wise bilinear
attention distributions to capture the 2^ nd order interactions between the in
put single-modal or multi-modal features. Higher and even infinity order feature
interactions are readily modeled through stacking multiple X-Linear attention b
locks and equipping the block with Exponential Linear Unit (ELU) in a parameter-
free fashion, respectively. Furthermore, we present X-Linear Attention Networks
(dubbed as X-LAN) that novelly integrates X-Linear attention block(s) into image
encoder and sentence decoder of image captioning model to leverage higher order
intra- and inter-modal interactions. The experiments on COCO benchmark demonstr
ate that our X-LAN obtains to-date the best published CIDEr performance of 132.0
% on COCO Karpathy test split. When further endowing Transformer with X-Linear a
ttention blocks, CIDEr is boosted up to 132.8%. Source code is available at http
s://github.com/Panda-Peter/image-captioning.
****

You2Me: Inferring Body Pose in Egocentric Video via First and Second Person Inte
ractions
Evonne Ng, Donglai Xiang, Hanbyul Joo, Kristen Grauman; Proceedings of the IE

EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9890-9900

The body pose of a person wearing a camera is of great interest for applications in augmented reality, healthcare, and robotics, yet much of the person's body is out of view for a typical wearable camera. We propose a learning-based approach to estimate the camera wearer's 3D body pose from egocentric video sequences. Our key insight is to leverage interactions with another person---whose body pose we can directly observe---as a signal inherently linked to the body pose of the first-person subject. We show that since interactions between individuals often induce a well-ordered series of back-and-forth responses, it is possible to learn a temporal model of the interlinked poses even though one party is largely out of view. We demonstrate our idea on a variety of domains with dyadic interaction and show the substantial impact on egocentric body pose estimation, which improves the state of the art.

****

Estimating Low-Rank Region Likelihood Maps

Gabriela Csurka, Zoltan Kato, Andor Juhasz, Martin Humenberger; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13776-13785

Low-rank regions capture geometrically meaningful structures in an image which encompass typical local features such as edges, corners and all kinds of regular, symmetric, often repetitive patterns, that are commonly found in man-made environment. While such patterns are challenging current state-of-the-art feature correspondence methods, the recovered homography of a low-rank texture readily provides 3D structure with respect to a 3D plane, without any prior knowledge of the visual information on that plane. However, the automatic and efficient detection of the broad class of low-rank regions is unsolved. Herein, we propose a novel self-supervised low-rank region detection deep network that predicts a low-rank likelihood map from an image. The evaluation of our method on real-world datasets shows not only that it reliably predicts low-rank regions in the image similarly to our baseline method, but thanks to the data augmentations used in the training phase it generalizes well to difficult cases (e.g. day/night lighting, low contrast, underexposure) where the baseline prediction fails.

****

ABCNet: Real-Time Scene Text Spotting With Adaptive Bezier-Curve Network

Yuliang Liu, Hao Chen, Chunhua Shen, Tong He, Lianwen Jin, Liangwei Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9809-9818

Scene text detection and recognition has received increasing research attention. Existing methods can be roughly categorized into two groups: character-based and segmentation-based. These methods either are costly for character annotation or need to maintain a complex pipeline, which is often not suitable for real-time applications. Here we address the problem by proposing the Adaptive Bezier-Curve Network (\BeCan). Our contributions are three-fold: 1) For the first time, we adaptively fit oriented or curved text by a parameterized Bezier curve. 2) We design a novel BezierAlign layer for extracting accurate convolution features of a text instance with arbitrary shapes, significantly improving the precision compared with previous methods. 3) Compared with standard bounding box detection, our Bezier curve detection introduces negligible computation overhead, resulting in superiority of our method in both efficiency and accuracy. Experiments on oriented or curved benchmark datasets, namely Total-Text and CTW1500, demonstrate that \BeCan achieves state-of-the-art accuracy, meanwhile significantly improving the speed. In particular, on Total-Text, our real-time version is over 10 times faster than recent state-of-the-art methods with a competitive recognition accuracy. Code is available at https://git.io/AdelaiDet.

****

FeatureFlow: Robust Video Interpolation via Structure-to-Texture Generation

Shurui Gui, Chaoyue Wang, Qihua Chen, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14004-14013

Video interpolation aims to synthesize non-existent frames between two consecuti ve frames. Although existing optical flow based methods have achieved promising results, they still face great challenges in dealing with the interpolation of c omplicated dynamic scenes, which include occlusion, blur or abrupt brightness ch ange. This is mainly because these cases may break the basic assumptions of the optical flow estimation (i.e. smoothness, consistency). In this work, we devised a novel structure-to-texture generation framework which splits the video interp olation task into two stages: structure-guided interpolation and texture refinem ent. In the first stage, deep structure-aware features are employed to predict f eature flows from two consecutive frames to their intermediate result, and furth er generate the structure image of the intermediate frame. In the second stage, based on the generated coarse result, a Frame Texture Compensator is trained to fill in detailed textures. To the best of our knowledge, this is the first work that attempts to directly generate the intermediate frame through blending deep features. Experiments on both the benchmark datasets and challenging occlusion c ases demonstrate the superiority of the proposed framework over the state-of-the -art methods. Codes are available on https://github.com/CM-BF/FeatureFlow.

****

Attention-Guided Hierarchical Structure Aggregation for Image Matting

Yu Qiao, Yuhao Liu, Xin Yang, Dongsheng Zhou, Mingliang Xu, Qiang Zhang, X iaopeng Wei; Proceedings of the IEEE/CVF Conference on Computer Vision and Patte rn Recognition (CVPR), 2020, pp. 13676-13685

Existing deep learning based matting algorithms primarily resort to high-level s emantic features to improve the overall structure of alpha mattes. However, we a rgue that advanced semantics extracted from CNNs contribute unequally for alpha perception and we are supposed to reconcile advanced semantic information with l ow-level appearance cues to refine the foreground details. In this paper, we pro pose an end-to-end Hierarchical Attention Matting Network (HAttMatting), which c an predict the better structure of alpha mattes from single RGB images without a dditional input. Specifically, we employ spatial and channel-wise attention to i ntegrate appearance cues and pyramidal features in a novel fashion. This blended attention mechanism can perceive alpha mattes from refined boundaries and adapt ive semantics. We also introduce a hybrid loss function fusing Structural SIMila rity (SSIM), Mean Square Error (MSE) and Adversarial loss to guide the network t o further improve the overall foreground structure. Besides, we construct a larg e-scale image matting dataset comprised of 59,600 training images and 1000 test images (total 646 distinct foreground alpha mattes), which can further improve t he robustness of our hierarchical structure aggregation model. Extensive experim ents demonstrate that the proposed HAttMatting can capture sophisticated foregro und structure and achieve state-of-the-art performance with single RGB images as input.

****

SharinGAN: Combining Synthetic and Real Data for Unsupervised Geometry Estimatio n

Koutilya PNVR, Hao Zhou, David Jacobs; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13974-13983

We propose a novel method for combining synthetic and real images when training networks to determine geometric information from a single image. We suggest a me thod for mapping both image types into a single, shared domain. This is connecte d to a primary network for end-to-end training. Ideally, this results in images from two domains that present shared information to the primary network. Our exp eriments demonstrate significant improvements over the state-of-the-art in two i mportant domains, surface normal estimation of human faces and monocular depth e stimation for outdoor scenes, both in an unsupervised setting.

****

Large-Scale Object Detection in the Wild From Imbalanced Multi-Labels

Junran Peng, Xingyuan Bu, Ming Sun, Zhaoxiang Zhang, Tieniu Tan, Junjie Yan ; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit ion (CVPR), 2020, pp. 9709-9718

Training with more data has always been the most stable and effective way of imp

roving performance in deep learn-ing era. As the largest object detection datase t so far, OpenImages brings great opportunities and challenges for object detect ion in general and sophisticated scenarios. However, owing to its semi-automatic collecting and labeling pipeline to deal with the huge data scale, Open Images dataset suffers from label-related problems that objects may explicitly or impli citly have multiple labels and the label distribution is extremely imbalanced. I n this work, we quantitatively analyze these label problems and provide a simple but effective solution. We design a concurrent softmax to handle the multi-labe l problems in object detection and propose a soft-sampling methods with hybrid t raining scheduler to deal with the label imbalance. Overall, our method yields a dramatic improvement of 3.34 points, leading to the best single model with 60.9 0 mAP on the public object detection test set of Open Images. And our ensembling result achieves 67.17mAP, which is 4.29 points higher than the first place meth od last year.

****

AugFPN: Improving Multi-Scale Feature Learning for Object Detection

Chaoxu Guo, Bin Fan, Qian Zhang, Shiming Xiang, Chunhong Pan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020 , pp. 12595-12604

Current state-of-the-art detectors typically exploit feature pyramid to detect o bjects at different scales. Among them, FPN is one of the representative works t hat build a feature pyramid by multi-scale features summation. However, the desi gn defects behind prevent the multi-scale features from being fully exploited. I n this paper, we begin by first analyzing the design defects of feature pyramid in FPN, and then introduce a new feature pyramid architecture named AugFPN to ad dress these problems. Specifically, AugFPN consists of three components: Consist ent Supervision, Residual Feature Augmentation, and Soft RoI Selection. AugFPN n arrows the semantic gaps between features of different scales before feature fus ion through Consistent Supervision. In feature fusion, ratio-invariant context i nformation is extracted by Residual Feature Augmentation to reduce the informati on loss of feature map at the highest pyramid level. Finally, Soft RoI Selection is employed to learn a better RoI feature adaptively after feature fusion. By r eplacing FPN with AugFPN in Faster R-CNN, our models achieve 2.3 and 1.6 points higher Average Precision (AP) when using ResNet50 and MobileNet-v2 as backbone r espectively. Furthermore, AugFPN improves RetinaNet by 1.6 points AP and FCOS by 0.9 points AP when using ResNet50 as backbone. Codes are available on https://g ithub.com/Gus-Guo/AugFPN.

****

Deep Residual Flow for Out of Distribution Detection

Ev Zisselman, Aviv Tamar; Proceedings of the IEEE/CVF Conference on Computer Vi sion and Pattern Recognition (CVPR), 2020, pp. 13994-14003

The effective application of neural networks in the real-world relies on profici ently detecting out-of-distribution examples. Contemporary methods seek to model the distribution of feature activations in the training data for adequately dis tinguishing abnormalities, and the state-of-the-art method uses Gaussian distrib ution models. In this work, we present a novel approach that improves upon the s tate-of-the-art by leveraging an expressive density model based on normalizing f lows. We introduce the residual flow, a novel flow architecture that learns the residual distribution from a base Gaussian distribution. Our model is general, a nd can be applied to any data that is approximately Gaussian. For out of distrib ution detection in image datasets, our approach provides a principled improvemen t over the state-of-the-art. Specifically, we demonstrate the effectiveness of o ur method in ResNet and DenseNet architectures trained on various image datasets . For example, on a ResNet trained on CIFAR-100 and evaluated on detection of ou t-of-distribution samples from the ImageNet dataset, holding the true positive r ate (TPR) at 95%, we improve the true negative rate (TNR) from 56.7% (current st ate of-the-art) to 77.5% (ours).

****

Don't Judge an Object by Its Context: Learning to Overcome Contextual Bias

Krishna Kumar Singh, Dhruv Mahajan, Kristen Grauman, Yong Jae Lee, Matt Feis

zli, Deepti Ghadiyaram; Proceedings of the IEEE/CVF Conference on Computer Visi on and Pattern Recognition (CVPR), 2020, pp. 11070-11078
Existing models often leverage co-occurrences between objects and their context to improve recognition accuracy. However, strongly relying on context risks a mo del's generalizability, especially when typical co-occurrence patterns are absen t. This work focuses on addressing such contextual biases to improve the robustn ess of the learnt feature representations. Our goal is to accurately recognize a category in the absence of its context, without compromising on performance whe n it co-occurs with context. Our key idea is to decorrelate feature representati ons of a category from its co-occurring context. We achieve this by learning a f eature subspace that explicitly represents categories occurring in the absence o f context along side a joint feature subspace that represents both categories an d context. Our very simple yet effective method is extensible to two multi-label tasks -- object and attribute classification. On 4 challenging datasets, we dem onstrate the effectiveness of our method in reducing contextual bias.
****

SEED: Semantics Enhanced Encoder-Decoder Framework for Scene Text Recognition
Zhi Qiao, Yu Zhou, Dongbao Yang, Yucan Zhou, Weiping Wang; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, p p. 13528-13537
Scene text recognition is a hot research topic in computer vision. Recently, man y recognition methods based on the encoder-decoder framework have been proposed, and they can handle scene texts of perspective distortion and curve shape. Neve rtheless, they still face lots of challenges like image blur, uneven illuminatio n, and incomplete characters. We argue that most encoder-decoder methods are bas ed on local visual features without explicit global semantic information. In thi s work, we propose a semantics enhanced encoder-decoder framework to robustly re cognize low-quality scene texts. The semantic information is used both in the en coder module for supervision and in the decoder module for initializing. In part icular, the state-of-the-art ASTER method is integrated into the proposed framew ork as an exemplar. Extensive experiments demonstrate that the proposed framewor k is more robust for low-quality text images, and achieves state-of-the-art resu lts on several benchmark datasets. The source code will be available.
****

Iterative Answer Prediction With Pointer-Augmented Multimodal Transformers for T extVQA
Ronghang Hu, Amanpreet Singh, Trevor Darrell, Marcus Rohrbach; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020 , pp. 9992-10002
Many visual scenes contain text that carries crucial information, and it is thus essential to understand text in images for downstream reasoning tasks. For exam ple, a deep water label on a warning sign warns people about the danger in the s cene. Recent work has explored the TextVQA task that requires reading and unders tanding text in images to answer a question. However, existing approaches for Te xtVQA are mostly based on custom pairwise fusion mechanisms between a pair of tw o modalities and are restricted to a single prediction step by casting TextVQA a s a classification task. In this work, we propose a novel model for the TextVQA task based on a multimodal transformer architecture accompanied by a rich repres entation for text in images. Our model naturally fuses different modalities homo geneously by embedding them into a common semantic space where self-attention is applied to model inter- and intra- modality context. Furthermore, it enables it erative answer decoding with a dynamic pointer network, allowing the model to fo rm an answer through multi-step prediction instead of one-step classification. O ur model outperforms existing approaches on three benchmark datasets for the Tex tVQA task by a large margin.
****

iTAML: An Incremental Task-Agnostic Meta-learning Approach
Jathushan Rajasegaran, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Mubar ak Shah; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern R ecognition (CVPR), 2020, pp. 13588-13597

Humans can continuously learn new knowledge as their experience grows. In contrast, previous learning in deep neural networks can quickly fade out when they are trained on a new task. In this paper, we hypothesize this problem can be avoided by learning a set of generalized parameters, that are neither specific to old nor new tasks. In this pursuit, we introduce a novel meta-learning approach that seeks to maintain an equilibrium between all the encountered tasks. This is ensured by a new meta-update rule which avoids catastrophic forgetting. In comparison to previous meta-learning techniques, our approach is task-agnostic. When presented with a continuum of data, our model automatically identifies the task and quickly adapts to it with just a single update. We perform extensive experiments on five datasets in a class-incremental setting, leading to significant improvements over the state of the art methods (e.g., a 21.3% boost on CIFAR100 with 10 incremental tasks). Specifically, on large-scale datasets that generally prove difficult cases for incremental learning, our approach delivers absolute gains as high as 19.1% and 7.4% on ImageNet and MS-Celeb datasets, respectively.
****

Open Compound Domain Adaptation
Ziwei Liu, Zhongqi Miao, Xingang Pan, Xiaohang Zhan, Dahua Lin, Stella X. Yu, Boqing Gong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12406-12415
A typical domain adaptation approach is to adapt models trained on the annotated data in a source domain (e.g., sunny weather) for achieving high performance on the test data in a target domain (e.g., rainy weather). Whether the target contains a single homogeneous domain or multiple heterogeneous domains, existing works always assume that there exist clear distinctions between the domains, which is often not true in practice (e.g., changes in weather). We study an open compound domain adaptation (OCDA) problem, in which the target is a compound of multiple homogeneous domains without domain labels, reflecting realistic data collection from mixed and novel situations. We propose a new approach based on two technical insights into OCDA: 1) a curriculum domain adaptation strategy to bootstrap generalization across domains in a data-driven self-organizing fashion and 2) a memory module to increase the model's agility towards novel domains. Our experiments on digit classification, facial expression recognition, semantic segmentation, and reinforcement learning demonstrate the effectiveness of our approach.
****

GrappaNet: Combining Parallel Imaging With Deep Learning for Multi-Coil MRI Reconstruction
Anuroop Sriram, Jure Zbontar, Tullie Murrell, C. Lawrence Zitnick, Aaron Defazio, Daniel K. Sodickson; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14315-14322
Magnetic Resonance Image (MRI) acquisition is an inherently slow process which has spurred the development of two different acceleration methods: acquiring multiple correlated samples simultaneously (parallel imaging) and acquiring fewer samples than necessary for traditional signal processing methods (compressed sensing). Both methods provide complementary approaches to accelerating MRI acquisition. In this paper, we present a novel method to integrate traditional parallel imaging methods into deep neural networks that is able to generate high quality reconstructions even for high acceleration factors. The proposed method, called GrappaNet, performs progressive reconstruction by first mapping the reconstruction problem to a simpler one that can be solved by a traditional parallel imaging methods using a neural network, followed by an application of a parallel imaging method, and finally fine-tuning the output with another neural network. The entire network can be trained end-to-end. We present experimental results on the recently released fastMRI dataset and show that GrappaNet can generate higher quality reconstructions than competing methods for both 4x and 8x acceleration.
****

FBNetV2: Differentiable Neural Architecture Search for Spatial and Channel Dimensions
Alvin Wan, Xiaoliang Dai, Peizhao Zhang, Zijian He, Yuandong Tian, Saining Xie, Bichen Wu, Matthew Yu, Tao Xu, Kan Chen, Peter Vajda, Joseph E. Gonza

lez; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog nition (CVPR), 2020, pp. 12965-12974
Differentiable Neural Architecture Search (DNAS) has demonstrated great success in designing state-of-the-art, efficient neural networks. However, DARTS-based D NAS's search space is small when compared to other search methods', since all ca ndidate network layers must be explicitly instantiated in memory. To address thi s bottleneck, we propose a memory and computationally efficient DNAS variant: DM askingNAS. This algorithm expands the search space by up to 10^14x over conventi onal DNAS, supporting searches over spatial and channel dimensions that are othe rwise prohibitively expensive: input resolution and number of filters. We propos e a masking mechanism for feature map reuse, so that memory and computational co sts stay nearly constant as the search space expands. Furthermore, we employ eff ective shape propagation to maximize per-FLOP or per-parameter accuracy. The sea rched FBNetV2s yield state-of-the-art performance when compared with all previou s architectures. With up to 421x less search cost, DMaskingNAS finds models with 0.9% higher accuracy, 15% fewer FLOPs than MobileNetV3-Small; and with similar accuracy but 20% fewer FLOPs than Efficient-B0. Furthermore, our FBNetV2 outperf orms MobileNetV3 by 2.6% in accuracy, with equivalent model size. FBNetV2 models are open-sourced at https://github.com/facebookresearch/mobile-vision.
****
Webly Supervised Knowledge Embedding Model for Visual Reasoning
Wenbo Zheng, Lan Yan, Chao Gou, Fei-Yue Wang; Proceedings of the IEEE/CVF Con ference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12445-12454
Visual reasoning between visual image and natural language description is a long -standing challenge in computer vision. While recent approaches offer a great pr omise by compositionality or relational computing, most of them are oppressed by the challenge of training with datasets containing only a limited number of ima ges with ground-truth texts. Besides, it is extremely time-consuming and difficu lt to build a larger dataset by annotating millions of images with text descript ions that may very likely lead to a biased model. Inspired by the majority succe ss of webly supervised learning, we utilize readily-available web images with it s noisy annotations for learning a robust representation. Our key idea is to pre sume on web images and corresponding tags along with fully annotated datasets in learning with knowledge embedding. We present a two-stage approach for the task that can augment knowledge through an effective embedding model with weakly sup ervised web data. This approach learns not only knowledge-based embeddings deriv ed from key-value memory networks to make joint and full use of textual and visu al information but also exploits the knowledge to improve the performance with k nowledge-based representation learning for applying other general reasoning task s. Experimental results on two benchmarks show that the proposed approach signif icantly improves performance compared with the state-of-the-art methods and guar antees the robustness of our model against visual reasoning tasks and other reas oning tasks.
****
Scale-Equalizing Pyramid Convolution for Object Detection
Xinjiang Wang, Shilong Zhang, Zhuoran Yu, Litong Feng, Wayne Zhang; Proceedi ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 2020, pp. 13359-13368
Feature pyramid has been an efficient method to extract features at different sc ales. Development over this method mainly focuses on aggregating contextual info rmation at different levels while seldom touching the inter-level correlation in the feature pyramid. Early computer vision methods extracted scale-invariant fe atures by locating the feature extrema in both spatial and scale dimension. Insp ired by this, a convolution across the pyramid level is proposed in this study, which is termed pyramid convolution and is a modified 3-D convolution. Stacked p yramid convolutions directly extract 3-D (scale and spatial) features and outper forms other meticulously designed feature fusion modules. Based on the viewpoint of 3-D convolution, an integrated batch normalization that collects statistics from the whole feature pyramid is naturally inserted after the pyramid convoluti on. Furthermore, we also show that the naive pyramid convolution, together with

the design of RetinaNet head, actually best applies for extracting features from a Gaussian pyramid, whose properties can hardly be satisfied by a feature pyramid. In order to alleviate this discrepancy, we build a scale-equalizing pyramid convolution (SEPC) that aligns the shared pyramid convolution kernel only at high-level feature maps. Being computationally efficient and compatible with the head design of most single-stage object detectors, the SEPC module brings significant performance improvement (>4AP increase on MS-COCO2017 dataset) in state-of-the-art one-stage object detectors, and a light version of SEPC also has 3.5AP gain with only around 7% inference time increase. The pyramid convolution also functions well as a stand-alone module in two-stage object detectors and is able to improve the performance by 2AP. The source code can be found at https://github.com/jshilong/SEPC.

****

Learning Selective Self-Mutual Attention for RGB-D Saliency Detection

Nian Liu, Ni Zhang, Junwei Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13756-13765

Saliency detection on RGB-D images is receiving more and more research interests recently. Previous models adopt the early fusion or the result fusion scheme to fuse the input RGB and depth data or their saliency maps, which incur the problem of distribution gap or information loss. Some other models use the feature fusion scheme but are limited by the linear feature fusion methods. In this paper, we propose to fuse attention learned in both modalities. Inspired by the Non-local model, we integrate the self-attention and each other's attention to propagate long-range contextual dependencies, thus incorporating multi-modal information to learn attention and propagate contexts more accurately. Considering the reliability of the other modality's attention, we further propose a selection attention to weight the newly added attention term. We embed the proposed attention module in a two-stream CNN for RGB-D saliency detection. Furthermore, we also propose a residual fusion module to fuse the depth decoder features into the RGB stream. Experimental results on seven benchmark datasets demonstrate the effectiveness of the proposed model components and our final saliency model. Our code and saliency maps are available at https://github.com/nnizhang/S2MA.

****

Prior Guided GAN Based Semantic Inpainting

Avisek Lahiri, Arnav Kumar Jain, Sanskar Agrawal, Pabitra Mitra, Prabir Kumar Biswas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13696-13705

Contemporary deep learning based semantic inpainting can be approached from two directions. First, and the more explored, approach is to train an offline deep regression network over the masked pixels with an additional refinement by adversarial training. This approach requires a single feed-forward pass for inpainting at inference. Another promising, yet unexplored approach is to first train a generative model to map a latent prior distribution to natural image manifold and during inference time search for the best-matching prior to reconstruct the signal. The primary aversion towards the latter genre is due to its inference time iterative optimization and difficulty to scale to higher resolution. In this paper, going against the general trend, we focus on the second paradigm of inpainting and address both of its mentioned problems. Most importantly, we learn a data driven parametric network to directly predict a matching prior for a given masked image. This converts an iterative paradigm to a single feed forward inference pipeline with around 800X speedup. We also regularize our network with structural prior (computed from the masked image itself) which helps in better preservation of pose and size of the object to be inpainted. Moreover, to extend our model for sequence reconstruction, we propose a recurrent net based grouped latent prior learning. Finally, we leverage recent advancements in high resolution GAN training to scale our inpainting network to 256X256. Experiments (spanning across resolutions from 64X64 to 256X256) conducted on SVHN, Standford Cars, CelebA, CelebA-HQ and ImageNet image datasets, and FaceForensics video datasets reveal that we consistently improve upon contemporary benchmarks from both schools of approaches.

****
DeFeat-Net: General Monocular Depth via Simultaneous Unsupervised Representation Learning
Jaime Spencer, Richard Bowden, Simon Hadfield; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14402-14413
In the current monocular depth research, the dominant approach is to employ unsupervised training on large datasets, driven by warped photometric consistency. Such approaches lack robustness and are unable to generalize to challenging domains such as nighttime scenes or adverse weather conditions where assumptions about photometric consistency break down. We propose DeFeat-Net (Depth & Feature network), an approach to simultaneously learn a cross-domain dense feature representation, alongside a robust depth-estimation framework based on warped feature consistency. The resulting feature representation is learned in an unsupervised manner with no explicit ground-truth correspondences required. We show that within a single domain, our technique is comparable to both the current state of the art in monocular depth estimation and supervised feature representation learning. However, by simultaneously learning features, depth and motion, our technique is able to generalize to challenging domains, allowing DeFeat-Net to outperform the current state-of-the-art with around 10% reduction in all error measures on more challenging sequences such as nighttime driving.
****
Regularizing Class-Wise Predictions via Self-Knowledge Distillation
Sukmin Yun, Jongjin Park, Kimin Lee, Jinwoo Shin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13876-13885
Deep neural networks with millions of parameters may suffer from poor generalization due to overfitting. To mitigate the issue, we propose a new regularization method that penalizes the predictive distribution between similar samples. In particular, we distill the predictive distribution between different samples of the same label during training. This results in regularizing the dark knowledge (i.e., the knowledge on wrong predictions) of a single network (i.e., a self-knowledge distillation) by forcing it to produce more meaningful and consistent predictions in a class-wise manner. Consequently, it mitigates overconfident predictions and reduces intra-class variations. Our experimental results on various image classification tasks demonstrate that the simple yet powerful method can significantly improve not only the generalization ability but also the calibration performance of modern convolutional neural networks.
****
Modality Shifting Attention Network for Multi-Modal Video Question Answering
Junyeong Kim, Minuk Ma, Trung Pham, Kyungsu Kim, Chang D. Yoo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10106-10115
This paper considers a network referred to as Modality Shifting Attention Network (MSAN) for Multimodal Video Question Answering (MVQA) task. MSAN decomposes the task into two sub-tasks: (1) localization of temporal moment relevant to the question, and (2) accurate prediction of the answer based on the localized moment. The modality required for temporal localization may be different from that for answer prediction, and this ability to shift modality is essential for performing the task. To this end, MSAN is based on (1) the moment proposal network (MPN) that attempts to locate the most appropriate temporal moment from each of the modalities, and also on (2) the heterogeneous reasoning network (HRN) that predicts the answer using an attention mechanism on both modalities. MSAN is able to place importance weight on the two modalities for each sub-task using a component referred to as Modality Importance Modulation (MIM). Experimental results show that MSAN outperforms previous state-of-the-art by achieving 71.13% test accuracy on TVQA benchmark dataset. Extensive ablation studies and qualitative analysis are conducted to validate various components of the network.
****
Vision-Language Navigation With Self-Supervised Auxiliary Reasoning Tasks
Fengda Zhu, Yi Zhu, Xiaojun Chang, Xiaodan Liang; Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10012-10022

Vision-Language Navigation (VLN) is a task where an agent learns to navigate following a natural language instruction. The key to this task is to perceive both the visual scene and natural language sequentially. Conventional approaches fully exploit vision and language features in cross-modal grounding. However, the VLN task remains challenging, since previous works have implicitly neglected the rich semantic information contained in environments (such as navigation graphs or sub-trajectory semantics). In this paper, we introduce Auxiliary Reasoning Navigation (AuxRN), a framework with four self-supervised auxiliary reasoning tasks to exploit the additional training signals derived from these semantic information. The auxiliary tasks have four reasoning objectives: explaining the previous actions, evaluating the trajectory consistency, estimating the progress and predict the next direction. As a result, these additional training signals help the agent to acquire knowledge of semantic representations in order to reason about its activities and build a thorough perception of environments. Our experiments demonstrate that auxiliary reasoning tasks improve both the performance of the main task and the model generalizability by a large margin. We further demonstrate empirically that an agent trained with self-supervised auxiliary reasoning tasks substantially outperforms the previous state-of-the-art method, being the best existing approach on the standard benchmark.

****

Hypergraph Attention Networks for Multimodal Learning

Eun-Sol Kim, Woo Young Kang, Kyoung-Woon On, Yu-Jung Heo, Byoung-Tak Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14581-14590

One of the fundamental problems that arise in multimodal learning tasks is the disparity of information levels between different modalities. To resolve this problem, we propose Hypergraph Attention Networks (HANs), which define a common semantic space among the modalities with symbolic graphs and extract a joint representation of the modalities based on a co-attention map constructed in the semantic space. HANs follow the process: constructing the common semantic space with symbolic graphs of each modality, matching the semantics between sub-structures of the symbolic graphs, constructing co-attention maps between the graphs in the semantic space, and integrating the multimodal inputs using the co-attention maps to get the final joint representation. From the qualitative analysis with two Visual Question and Answering datasets, we discover that 1) the alignment of the information levels between the modalities is important, and 2) the symbolic graphs are very powerful ways to represent the information of the low-level signals in alignment. Moreover, HANs dramatically improve the state-of-the-art accuracy on the GQA dataset from 54.6% to 61.88% only using the symbolic information in quantitatively.

****

RL-CycleGAN: Reinforcement Learning Aware Simulation-to-Real

Kanishka Rao, Chris Harris, Alex Irpan, Sergey Levine, Julian Ibarz, Mohi Khansari; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11157-11166

Deep neural network based reinforcement learning (RL) can learn appropriate visual representations for complex tasks like vision-based robotic grasping without the need for manually engineering or prior learning a perception system. However, data for RL is collected via running an agent in the desired environment, and for applications like robotics, running a robot in the real world may be extremely costly and time consuming. Simulated training offers an appealing alternative, but ensuring that policies trained in simulation can transfer effectively into the real world requires additional machinery. Simulations may not match reality, and typically bridging the simulation-to-reality gap requires domain knowledge and task-specific engineering. We can automate this process by employing generative models to translate simulated images into realistic ones. However, this sort of translation is typically task-agnostic, in that the translated images may not preserve all features that are relevant to the task. In this paper, we introd

uce the RL-scene consistency loss for image translation, which ensures that the translation operation is invariant with respect to the Q-values associated with the image. This allows us to learn a task-aware translation. Incorporating this loss into unsupervised domain translation, we obtain the RL-CycleGAN, a new approach for simulation-to-real-world transfer for reinforcement learning. In evaluations of RL-CycleGAN on two vision-based robotics grasping tasks, we show that RL-CycleGAN offers a substantial improvement over a number of prior methods for sim-to-real transfer, attaining excellent real-world performance with only a modest number of real-world observations.
****

Video Instance Segmentation Tracking With a Modified VAE Architecture
Chung-Ching Lin, Ying Hung, Rogerio Feris, Linglin He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13147-13157
We propose a modified variational autoencoder (VAE) architecture built on top of Mask R-CNN for instance-level video segmentation and tracking. The method builds a shared encoder and three parallel decoders, yielding three disjoint branches for predictions of future frames, object detection boxes, and instance segmentation masks. To effectively solve multiple learning tasks, we introduce a Gaussian Process model to enhance the statistical representation of VAE by relaxing the prior strong independent and identically distributed (iid) assumption of conventional VAEs and allowing potential correlations among extracted latent variables. The network learns embedded spatial interdependence and motion continuity in video data and creates a representation that is effective to produce high-quality segmentation masks and track multiple instances in diverse and unstructured videos. Evaluation on a variety of recently introduced datasets shows that our model outperforms previous methods and achieves the new best in class performance.
****

High-Dimensional Convolutional Networks for Geometric Pattern Recognition
Christopher Choy, Junha Lee, Rene Ranftl, Jaesik Park, Vladlen Koltun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11227-11236
High-dimensional geometric patterns appear in many computer vision problems. In this work, we present high-dimensional convolutional networks for geometric pattern recognition problems that arise in 2D and 3D registration problems. We first propose high-dimensional convolutional networks from 4 to 32 dimensions and analyze the geometric pattern recognition capacity in high-dimensional linear regression problems. Next, we show that the 3D correspondences form hyper-surface in a 6-dimensional space and validate our network on 3D registration problems. Finally, we use image correspondences, which form a 4-dimensional hyper-conic section, and show that the high-dimensional convolutional networks are on par with many state-of-the-art multi-layered perceptrons.
****

AOWS: Adaptive and Optimal Network Width Search With Latency Constraints
Maxim Berman, Leonid Pishchulin, Ning Xu, Matthew B. Blaschko, Gerard Medioni; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11217-11226
Neural architecture search (NAS) approaches aim at automatically finding novel CNN architectures that fit computational constraints while maintaining a good performance on the target platform. We introduce a novel efficient one-shot NAS approach to optimally search for channel numbers, given latency constraints on a specific hardware. We first show that we can use a black-box approach to estimate a realistic latency model for a specific inference platform, without the need for low-level access to the inference computation. Then, we design a pairwise MRF to score any channel configuration and use dynamic programming to efficiently decode the best performing configuration, yielding an optimal solution for the network width search. Finally, we propose an adaptive channel configuration sampling scheme to gradually specialize the training phase to the target computational constraints. Experiments on ImageNet classification show that our approach can find networks fitting the resource constraints on different target platforms whil

e improving accuracy over the state-of-the-art efficient networks.
****

Attentive Weights Generation for Few Shot Learning via Information Maximization
Yiluan Guo, Ngai-Man Cheung; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13499-13508
Few shot image classification aims at learning a classifier from limited labeled data. Generating the classification weights has been applied in many meta-learning methods for few shot image classification due to its simplicity and effectiveness. In this work, we present Attentive Weights Generation for few shot learning via Information Maximization (AWGIM), which introduces two novel contributions: i) Mutual information maximization between generated weights and data within the task; this enables the generated weights to retain information of the task and the specific query sample. ii) Self-attention and cross-attention paths to encode the context of the task and individual queries. Both two contributions are shown to be very effective in extensive experiments. Overall, AWGIM is competitive with state-of-the-art. Code is available at https://github.com/Yiluan/AWGIM.
****

DeepLPF: Deep Local Parametric Filters for Image Enhancement
Sean Moran, Pierre Marza, Steven McDonagh, Sarah Parisot, Gregory Slabaugh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12826-12835
Digital artists often improve the aesthetic quality of digital photographs through manual retouching. Beyond global adjustments, professional image editing programs provide local adjustment tools operating on specific parts of an image. Options include parametric (graduated, radial filters) and unconstrained brush tools. These highly expressive tools enable a diverse set of local image enhancements. However, their use can be time consuming, and requires artistic capability. State-of-the-art automated image enhancement approaches typically focus on learning pixel-level or global enhancements. The former can be noisy and lack interpretability, while the latter can fail to capture fine-grained adjustments. In this paper, we introduce a novel approach to automatically enhance images using learned spatially local filters of three different types (Elliptical Filter, Graduated Filter, Polynomial Filter). We introduce a deep neural network, dubbed Deep Local Parametric Filters (DeepLPF), which regresses the parameters of these spatially localized filters that are then automatically applied to enhance the image. DeepLPF provides a natural form of model regularization and enables interpretable, intuitive adjustments that lead to visually pleasing results. We report on multiple benchmarks and show that DeepLPF produces state-of-the-art performance on two variants of the MIT-Adobe 5k dataset, often using a fraction of the parameters required for competing methods.
****

Joint Graph-Based Depth Refinement and Normal Estimation
Mattia Rossi, Mireille El Gheche, Andreas Kuhn, Pascal Frossard; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12154-12163
Depth estimation is an essential component in understanding the 3D geometry of a scene, with numerous applications in urban and indoor settings. These scenarios are characterized by a prevalence of human made structures, which in most of the cases are either inherently piece-wise planar or can be approximated as such. With these settings in mind, we devise a novel depth refinement framework that aims at recovering the underlying piece-wise planarity of those inverse depth maps associated to piece-wise planar scenes. We formulate this task as an optimization problem involving a data fidelity term, which minimizes the distance to the noisy and possibly incomplete input inverse depth map, as well as a regularization, which enforces a piece-wise planar solution. As for the regularization term, we model the inverse depth map pixels as the nodes of a weighted graph, with the weight of the edge between two pixels capturing the likelihood that they belong to the same plane in the scene. The proposed regularization fits a plane at each pixel automatically, avoiding any a priori estimation of the scene planes, and enforces that strongly connected pixels are assigned to the same plane. The re

sulting optimization problem is solved efficiently with the ADAM solver. Extensive tests show that our method leads to a significant improvement in depth refinement, both visually and numerically, with respect to state-of-the-art algorithms on the Middlebury, KITTI and ETH3D multi-view datasets.
****

Recognizing Objects From Any View With Object and Viewer-Centered Representations

Sainan Liu, Vincent Nguyen, Isaac Rehg, Zhuowen Tu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11784-11793

In this paper, we tackle an important task in computer vision: any view object recognition. In both training and testing, for each object instance, we are only given its 2D image viewed from an unknown angle. We propose a computational framework by designing object and viewer-centered neural networks (OVCNet) to recognize an object instance viewed from an arbitrary unknown angle. OVCNet consists of three branches that respectively implement object-centered, 3D viewer-centered, and in-plane viewer-centered recognition. We evaluate our proposed OVCNet using two metrics with unseen views from both seen and novel object instances. Experimental results demonstrate the advantages of OVCNet over classic 2D-image-based CNN classifiers, 3D-object (inferred from 2D image) classifiers, and competing multi-view based approaches. It gives rise to a viable and practical computing framework that combines both viewpoint-dependent and viewpoint-independent features for object recognition from any view.
****

Learning to Segment the Tail

Xinting Hu, Yi Jiang, Kaihua Tang, Jingyuan Chen, Chunyan Miao, Hanwang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14045-14054

Real-world visual recognition requires handling the extreme sample imbalance in large-scale long-tailed data. We propose a "divide&conquer" strategy for the challenging LVIS task: divide the whole data into balanced parts and then apply incremental learning to conquer each one. This derives a novel learning paradigm: class-incremental few-shot learning, which is especially effective for the challenge evolving over time: 1) the class imbalance among the old class knowledge review and 2) the few-shot data in new-class learning. We call our approach Learning to Segment the Tail (LST). In particular, we design an instance-level balanced replay scheme, which is a memory-efficient approximation to balance the instance-level samples from the old-class images. We also propose to use a meta-module for new-class learning, where the module parameters are shared across incremental phases, gaining the learning-to-learn knowledge incrementally, from the data-rich head to the data-poor tail. We empirically show that: at the expense of a little sacrifice of head-class forgetting, we can gain a significant 8.3% AP improvement for the tail classes with less than 10 instances, achieving an overall 2.0% AP boost for the whole 1,230 classes.
****

ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, Dieter Fox; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10740-10749

We present ALFRED (Action Learning From Realistic Environments and Directives), a benchmark for learning a mapping from natural language instructions and egocentric vision to sequences of actions for household tasks. ALFRED includes long, compositional tasks with non-reversible state changes to shrink the gap between research benchmarks and real-world applications. ALFRED consists of expert demonstrations in interactive visual environments for 25k natural language directives. These directives contain both high-level goals like "Rinse off a mug and place it in the coffee maker." and low-level language instructions like "Walk to the coffee maker on the right." ALFRED tasks are more complex in terms of sequence length, action space, and language than existing vision- and-language task datasets. We show that a baseline model based on recent embodied vision-and-language ta

sks performs poorly on ALFRED, suggesting that there is significant room for dev eloping innovative grounded visual language understanding models with this bench mark.
****

Robust Object Detection Under Occlusion With Context-Aware CompositionalNets
Angtian Wang, Yihong Sun, Adam Kortylewski, Alan L. Yuille; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, p p. 12645-12654
Detecting partially occluded objects is a difficult task. Our experimental resul ts show that deep learning approaches, such as Faster R-CNN, are not robust at o bject detection under occlusion. Compositional convolutional neural networks (Co mpositionalNets) have been shown to be robust at classifying occluded objects by explicitly representing the object as a composition of parts. In this work, we propose to overcome two limitations of CompositionalNets which will enable them to detect partially occluded objects: 1) CompositionalNets, as well as other DCN N architectures, do not explicitly separate the representation of the context fr om the object itself. Under strong object occlusion, the influence of the contex t is amplified which can have severe negative effects for detection at test time . In order to overcome this, we propose to segment the context during training v ia bounding box annotations. We then use the segmentation to learn a context-awa re compositionalNet that disentangles the representation of the context and the object. 2) We extend the part-based voting scheme in CompositionalNets to vote f or the corners of the object's bounding box, which enables the model to reliably estimate bounding boxes for partially occluded objects. Our extensive experimen ts show that our proposed model can detect objects robustly, increasing the dete ction performance of strongly occluded vehicles from PASCAL3D+ and MS-COCO by 41 % and 35% respectively in absolute performance relative to Faster R-CNN.
****

MnasFPN: Learning Latency-Aware Pyramid Architecture for Object Detection on Mob ile Devices
Bo Chen, Golnaz Ghiasi, Hanxiao Liu, Tsung-Yi Lin, Dmitry Kalenichenko, Har twig Adam, Quoc V. Le; Proceedings of the IEEE/CVF Conference on Computer Visio n and Pattern Recognition (CVPR), 2020, pp. 13607-13616
Despite the blooming success of architecture search for vision tasks in resource -constrained environments, the design of on-device object detection architecture s have mostly been manual. The few automated search efforts are either centered around non-mobile-friendly search spaces or not guided by on-device latency. We propose MnasFPN, a mobile-friendly search space for the detection head, and comb ine it with latency-aware architecture search to produce efficient object detect ion models. The learned MnasFPN head, when paired with MobileNetV2 body, outperf orms MobileNetV3+SSDLite by 1.8 mAP at similar latency on Pixel. It is both 1 mA P more accurate and 10% faster than NAS-FPNLite. Ablation studies show that the majority of the performance gain comes from innovations in the search space. Fur ther explorations reveal an interesting coupling between the search space design and the search algorithm, for which the complexity of MnasFPN search space is o pportune.
****

Cylindrical Convolutional Networks for Joint Object Detection and Viewpoint Esti mation
Sunghun Joung, Seungryong Kim, Hanjae Kim, Minsu Kim, Ig-Jae Kim, Junghyun Cho, Kwanghoon Sohn; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14163-14172
Existing techniques to encode spatial invariance within deep convolutional neura l networks only model 2D transformation fields. This does not account for the fa ct that objects in a 2D space are a projection of 3D ones, and thus they have li mited ability to severe object viewpoint changes. To overcome this limitation, w e introduce a learnable module, cylindrical convolutional networks (CCNs), that exploit cylindrical representation of a convolutional kernel defined in the 3D s pace. CCNs extract a view-specific feature through a view-specific convolutional kernel to predict object category scores at each viewpoint. With the view-speci

fic feature, we simultaneously determine objective category and viewpoints using the proposed sinusoidal soft-argmax module. Our experiments demonstrate the effectiveness of the cylindrical convolutional networks on joint object detection and viewpoint estimation.

****

Straight to the Point: Fast-Forwarding Videos via Reinforcement Learning Using Textual Data

Washington Ramos, Michel Silva, Edson Araujo, Leandro Soriano Marcolino, Erickson Nascimento; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10931-10940

The rapid increase in the amount of published visual data and the limited time of users bring the demand for processing untrimmed videos to produce shorter versions that convey the same information. Despite the remarkable progress that has been made by summarization methods, most of them can only select a few frames or skims, which creates visual gaps and breaks the video context. In this paper, we present a novel methodology based on a reinforcement learning formulation to accelerate instructional videos. Our approach can adaptively select frames that are not relevant to convey the information without creating gaps in the final video. Our agent is textually and visually oriented to select which frames to remove to shrink the input video. Additionally, we propose a novel network, called Visually-guided Document Attention Network (VDAN), able to generate a highly discriminative embedding space to represent both textual and visual data. Our experiments show that our method achieves the best performance in terms of F1 Score and coverage at the video segment level.

****

SPARE3D: A Dataset for SPAtial REasoning on Three-View Line Drawings

Wenyu Han, Siyuan Xiang, Chenhui Liu, Ruoyu Wang, Chen Feng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14690-14699

Spatial reasoning is an important component of human intelligence. We can imagine the shapes of 3D objects and reason about their spatial relations by merely looking at their three-view line drawings in 2D, with different levels of competence. Can deep networks be trained to perform spatial reasoning tasks? How can we measure their "spatial intelligence"? To answer these questions, we present the SPARE3D dataset. Based on cognitive science and psychometrics, SPARE3D contains three types of 2D-3D reasoning tasks on view consistency, camera pose, and shape generation, with increasing difficulty. We then design a method to automatically generate a large number of challenging questions with ground truth answers for each task. They are used to provide supervision for training our baseline models using state-of-the-art architectures like ResNet. Our experiments show that although convolutional networks have achieved superhuman performance in many visual learning tasks, their spatial reasoning performance in SPARE3D is almost equal to random guesses. We hope SPARE3D can stimulate new problem formulations and network designs for spatial reasoning to empower intelligent robots to operate effectively in the 3D world via 2D sensors.

****

Learning the Redundancy-Free Features for Generalized Zero-Shot Object Recognition

Zongyan Han, Zhenyong Fu, Jian Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12865-12874

Zero-shot object recognition or zero-shot learning aims to transfer the object recognition ability among the semantically related categories, such as fine-grained animal or bird species. However, the images of different fine-grained objects tend to merely exhibit subtle differences in appearance, which will severely deteriorate zero-shot object recognition. To reduce the superfluous information in the fine-grained objects, in this paper, we propose to learn the redundancy-free features for generalized zero-shot learning. We achieve our motivation by projecting the original visual features into a new (redundancy-free) feature space and then restricting the statistical dependence between these two feature spaces. Furthermore, we require the projected features to keep and even strengthen the

category relationship in the redundancy-free feature space. In this way, we can remove the redundant information from the visual features without losing the discriminative information. We extensively evaluate the performance on four benchmark datasets. The results show that our redundancy-free feature based generalized zero-shot learning (RFF-GZSL) approach can outperform the state-of-the-arts often by a large margin.
****

Instance-Aware, Context-Focused, and Memory-Efficient Weakly Supervised Object Detection
Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Yong Jae Lee, Alexander G. Schwing, Jan Kautz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10598-10607
Weakly supervised learning has emerged as a compelling tool for object detection by reducing the need for strong supervision during training. However, major challenges remain: (1) differentiation of object instances can be ambiguous; (2) detectors tend to focus on discriminative parts rather than entire objects; (3) without ground truth, object proposals have to be redundant for high recalls, causing significant memory consumption. Addressing these challenges is difficult, as it often requires to eliminate uncertainties and trivial solutions. To target these issues we develop an instance-aware and context-focused unified framework. It employs an instance-aware self-training algorithm and a learnable Concrete DropBlock while devising a memory-efficient sequential batch back-propagation. Our proposed method achieves state-of-the-art results on COCO (12.1% AP, 24.8% AP50), VOC 2007 (54.9% AP), and VOC 2012 (52.1% AP), improving baselines by great margins. In addition, the proposed method is the first to benchmark ResNet based models and weakly supervised video object detection. Refer to our project page for code, models, and more details: https://github.com/NVlabs/wetectron.
****

Differential Treatment for Stuff and Things: A Simple Unsupervised Domain Adaptation Method for Semantic Segmentation
Zhonghao Wang, Mo Yu, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wen-mei Hwu, Thomas S. Huang, Honghui Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12635-12644
We consider the problem of unsupervised domain adaptation for semantic segmentation by easing the domain shift between the source domain (synthetic data) and the target domain (real data) in this work. State-of-the-art approaches prove that performing semantic-level alignment is helpful in tackling the domain shift issue. Based on the observation that stuff categories usually share similar appearances across images of different domains while things (i.e. object instances) have much larger differences, we propose to improve the semantic-level alignment with different strategies for stuff regions and for things: 1) for the stuff categories, we generate feature representation for each class and conduct the alignment operation from the target domain to the source domain; 2) for the thing categories, we generate feature representation for each individual instance and encourage the instance in the target domain to align with the most similar one in the source domain. In this way, the individual differences within thing categories will also be considered to alleviate over-alignment. In addition to our proposed method, we further reveal the reason why the current adversarial loss is often unstable in minimizing the distribution discrepancy and show that our method can help ease this issue by minimizing the most similar stuff and instance features between the source and the target domains. We conduct extensive experiments in two unsupervised domain adaptation tasks, i.e. GTA5 - Cityscapes and SYNTHIA - Cityscapes, and achieve the new state-of-the-art segmentation accuracy.
****

One-Shot Adversarial Attacks on Visual Tracking With Dual Attention
Xuesong Chen, Xiyu Yan, Feng Zheng, Yong Jiang, Shu-Tao Xia, Yong Zhao, Rongrong Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10176-10185
Almost all adversarial attacks in computer vision are aimed at pre-known object categories, which could be offline trained for generating perturbations. But as

for visual object tracking, the tracked target categories are normally unknown i
n advance. However, the tracking algorithms also have potential risks of being a
ttacked, which could be maliciously used to fool the surveillance systems. Meanw
hile, it is still a challenging task that adversarial attacks on tracking since
it has the free-model tracked target. Therefore, to help draw more attention to
the potential risks, we study adversarial attacks on tracking algorithms. In thi
s paper, we propose a novel one-shot adversarial attack method to generate adver
sarial examples for free-model single object tracking, where merely adding sligh
t perturbations on the target patch in the initial frame causes state-of-the-art
 trackers to lose the target in subsequent frames. Specifically, the optimizatio
n objective of the proposed attack consists of two components and leverages the
dual attention mechanisms. The first component adopts a targeted attack strategy
 by optimizing the batch confidence loss with confidence attention while the sec
ond one applies a general perturbation strategy by optimizing the feature loss w
ith channel attention. Experimental results show that our approach can significa
ntly lower the accuracy of the most advanced Siamese network-based trackers on t
hree benchmarks.
****

Video Object Grounding Using Semantic Roles in Language Description
Arka Sadhu, Kan Chen, Ram Nevatia; Proceedings of the IEEE/CVF Conference on C
omputer Vision and Pattern Recognition (CVPR), 2020, pp. 10417-10427
We explore the task of Video Object Grounding (VOG), which grounds objects in vi
deos referred to in natural language descriptions. Previous methods apply image
grounding based algorithms to address VOG, fail to explore the object relation i
nformation and suffer from limited generalization. Here, we investigate the role
 of object relations in VOG and propose a novel framework VOGNet to encode multi
-modal object relations via self-attention with relative position encoding. To e
valuate VOGNet, we propose novel contrasting sampling methods to generate more c
hallenging grounding input samples, and construct a new dataset called ActivityN
et-SRL (ASRL) based on existing caption and grounding datasets. Experiments on A
SRL validate the need of encoding object relations in VOG, and our VOGNet outper
forms competitive baselines by a significant margin.
****

Context Prior for Scene Segmentation
Changqian Yu, Jingbo Wang, Changxin Gao, Gang Yu, Chunhua Shen, Nong Sang;
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio
n (CVPR), 2020, pp. 12416-12425
Recent works have widely explored the contextual dependencies to achieve more ac
curate segmentation results. However, most approaches rarely distinguish differe
nt types of contextual dependencies, which may pollute the scene understanding.
In this work, we directly supervise the feature aggregation to distinguish the i
ntra-class and interclass context clearly. Specifically, we develop a Context Pr
ior with the supervision of the Affinity Loss. Given an input image and correspo
nding ground truth, Affinity Loss constructs an ideal affinity map to supervise
the learning of Context Prior. The learned Context Prior extracts the pixels bel
onging to the same category, while the reversed prior focuses on the pixels of d
ifferent classes. Embedded into a conventional deep CNN, the proposed Context Pr
ior Layer can selectively capture the intra-class and inter-class contextual dep
endencies, leading to robust feature representation. To validate the effectivene
ss, we design an effective Context Prior Network (CPNet). Extensive quantitative
 and qualitative evaluations demonstrate that the proposed model performs favora
bly against state-of-the-art semantic segmentation approaches. More specifically
, our algorithm achieves 46.3% mIoU on ADE20K, 53.9% mIoU on PASCAL-Context, and
 81.3% mIoU on Cityscapes. Code is available at https://git.io/ContextPrior.
****

Binarizing MobileNet via Evolution-Based Searching
Hai Phan, Zechun Liu, Dang Huynh, Marios Savvides, Kwang-Ting Cheng, Zhiqia
ng Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern R
ecognition (CVPR), 2020, pp. 13420-13429
Binary Neural Networks (BNNs), known to be one among the effectively compact net

work architectures, have achieved great outcomes in the visual tasks. Designing efficient binary architectures is not trivial due to the binary nature of the network. In this paper, we propose a use of evolutionary search to facilitate the construction and training scheme when binarizing MobileNet, a compact network with separable depth-wise convolution. Being inspired by one-shot architecture search frameworks, we manipulate the idea of group convolution to design efficient 1-Bit Convolutional Neural Networks (CNNs), assuming an approximately optimal trade-off between computational cost and model accuracy. Our objective is to come up with a tiny yet efficient binary neural architecture by exploring the best candidates of the group convolution while optimizing the model performance in terms of complexity and latency. The approach is threefold. First, we modify and train strong baseline binary networks with a wide range of random group combinations at each convolutional layer. This set-up gives the binary neural networks a capability of preserving essential information through layers. Second, to find a good set of hyper-parameters for group convolutions we make use of the evolutionary search which leverages the exploration of efficient 1-bit models. Lastly, these binary models are trained from scratch in a usual manner to achieve the final binary model. Various experiments on ImageNet are conducted to show that following our construction guideline, the final model achieves 60.09% Top-1 accuracy and outperforms the state-of-the-art CI-BCNN with the same computational cost.

****

Adaptive Hierarchical Down-Sampling for Point Cloud Classification

Ehsan Nezhadarya, Ehsan Taghavi, Ryan Razani, Bingbing Liu, Jun Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12956-12964

Deterministic down-sampling of an unordered point cloud in a deep neural network has not been rigorously studied so far. Existing methods down-sample the points regardless of their importance for the network output and often address down-sampling the raw point cloud before processing. As a result, some important points in the point cloud may be removed, while less valuable points may be passed to next layers. In contrast, the proposed adaptive down-sampling method samples the points by taking into account the importance of each point, which varies according to application, task and training data. In this paper, we propose a novel deterministic, adaptive, permutation-invariant down-sampling layer, called Critical Points Layer (CPL), which learns to reduce the number of points in an unordered point cloud while retaining the important (critical) ones. Unlike most graph-based point cloud down-sampling methods that use k-NN to find the neighboring points, CPL is a global down-sampling method, rendering it computationally very efficient. The proposed layer can be used along with a graph-based point cloud convolution layer to form a convolutional neural network, dubbed CP-Net in this paper. We introduce a CP-Net for 3D object classification that achieves high accuracy for the ModelNet 40 dataset among point cloud-based methods, which validates the effectiveness of the CPL.

****

Learning Multi-View Camera Relocalization With Graph Neural Networks

Fei Xue, Xin Wu, Shaojun Cai, Junqiu Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11375-11384

We propose to construct a view graph to excavate the information of the whole given sequence for absolute camera pose estimation. Specifically, we harness GNNs to model the graph, allowing even non-consecutive frames to exchange information with each other. Rather than adopting the regular GNNs directly, we redefine the nodes, edges, and embedded functions to fit the relocalization task. Redesigned GNNs cooperate with CNNs in guiding knowledge propagation and feature extraction respectively to process multi-view high-dimension image features iteratively at different levels. Besides, a general graph-based loss function beyond constraints between consecutive views is employed for training the network in an end-to-end fashion. Extensive experiments conducted on both indoor and outdoor datasets demonstrate that our method outperforms previous approaches especially in large-scale and challenging scenarios.

****

Distortion Agnostic Deep Watermarking
Xiyang Luo, Ruohan Zhan, Huiwen Chang, Feng Yang, Peyman Milanfar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13548-13557
Watermarking is the process of embedding information into an image that can survive under distortions, while requiring the encoded image to have little or no perceptual difference with the original image. Recently, deep learning-based methods achieved impressive results in both visual quality and message payload under a wide variety of image distortions. However, these methods all require differentiable models for the image distortions at training time, and may generalize poorly to unknown distortions. This is undesirable since the types of distortions applied to watermarked images are usually unknown and non-differentiable. In this paper, we propose a new framework for distortion-agnostic watermarking, where the image distortion is not explicitly modeled during training. Instead, the robustness of our system comes from two sources: adversarial training and channel coding. Compared to training on a fixed set of distortions and noise levels, our method achieves comparable or better results on distortions available during training, and better performance overall on unknown distortions.
****
Disp R-CNN: Stereo 3D Object Detection via Shape Prior Guided Instance Disparity Estimation
Jiaming Sun, Linghao Chen, Yiming Xie, Siyu Zhang, Qinhong Jiang, Xiaowei Zhou, Hujun Bao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10548-10557
In this paper, we propose a novel system named Disp R-CNN for 3D object detection from stereo images. Many recent works solve this problem by first recovering a point cloud with disparity estimation and then apply a 3D detector. The disparity map is computed for the entire image, which is costly and fails to leverage category-specific prior. In contrast, we design an instance disparity estimation network (iDispNet) that predicts disparity only for pixels on objects of interest and learns a category-specific shape prior for more accurate disparity estimation. To address the challenge from scarcity of disparity annotation in training, we propose to use a statistical shape model to generate dense disparity pseudo-ground-truth without the need of LiDAR point clouds, which makes our system more widely applicable. Experiments on the KITTI dataset show that, even when LiDAR ground-truth is not available at training time, Disp R-CNN achieves competitive performance and outperforms previous state-of-the-art methods by 20% in terms of average precision. The code will be available at https://github.com/zju3dv/disprcnn.
****
Episode-Based Prototype Generating Network for Zero-Shot Learning
Yunlong Yu, Zhong Ji, Jungong Han, Zhongfei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14035-14044
We introduce a simple yet effective episode-based training framework for zero-shot learning (ZSL), where the learning system requires to recognize unseen classes given only the corresponding class semantics. During training, the model is trained within a collection of episodes, each of which is designed to simulate a zero-shot classification task. Through training multiple episodes, the model progressively accumulates ensemble experiences on predicting the mimetic unseen classes, which will generalize well on the real unseen classes. Based on this training framework, we propose a novel generative model that synthesizes visual prototypes conditioned on the class semantic prototypes. The proposed model aligns the visual-semantic interactions by formulating both the visual prototype generation and the class semantic inference into an adversarial framework paired with a parameter-economic Multi-modal Cross-Entropy Loss to capture the discriminative information. Extensive experiments on four datasets under both traditional ZSL and generalized ZSL tasks show that our model outperforms the state-of-the-art approaches by large margins.
****

Multi-Granularity Reference-Aided Attentive Feature Aggregation for Video-Based Person Re-Identification
Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Zhibo Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10407-10416
Video-based person re-identification (reID) aims at matching the same person across video clips. It is a challenging task due to the existence of redundancy among frames, newly revealed appearance, occlusion, and motion blurs. In this paper, we propose an attentive feature aggregation module, namely Multi-Granularity Reference-aided Attentive Feature Aggregation (MG-RAFA), to delicately aggregate spatio-temporal features into a discriminative video-level feature representation. In order to determine the contribution/importance of a spatial-temporal feature node, we propose to learn the attention from a global view with convolutional operations. Specifically, we stack its relations, i.e.no, pairwise correlations with respect to a representative set of reference feature nodes (S-RFNs) that represents global video information, together with the feature itself to infer the attention. Moreover, to exploit the semantics of different levels, we propose to learn multi-granularity attentions based on the relations captured at different granularities. Extensive ablation studies demonstrate the effectiveness of our attentive feature aggregation module MG-RAFA. Our framework achieves the state-of-the-art performance on three benchmark datasets.
****
Semi-Supervised Semantic Segmentation With Cross-Consistency Training
Yassine Ouali, Celine Hudelot, Myriam Tami; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12674-12684
In this paper, we present a novel cross-consistency based semi-supervised approach for semantic segmentation. Consistency training has proven to be a powerful semi-supervised learning framework for leveraging unlabeled data under the cluster assumption, in which the decision boundary should lie in low-density regions. In this work, we first observe that for semantic segmentation, the low-density regions are more apparent within the hidden representations than within the inputs. We thus propose cross-consistency training, where an invariance of the predictions is enforced over different perturbations applied to the outputs of the encoder. Concretely, a shared encoder and a main decoder are trained in a supervised manner using the available labeled examples. To leverage the unlabeled examples, we enforce a consistency between the main decoder predictions and those of the auxiliary decoders, taking as inputs different perturbed versions of the encoder's output, and consequently, improving the encoder's representations. The proposed method is simple and can easily be extended to use additional training signal, such as image-level labels or pixel-level labels across different domains. We perform an ablation study to tease apart the effectiveness of each component, and conduct extensive experiments to demonstrate that our method achieves state-of-the-art results in several datasets.
****
GaitPart: Temporal Part-Based Model for Gait Recognition
Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, Zhiqiang He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14225-14233
Gait recognition, applied to identify individual walking patterns in a long-distance, is one of the most promising video-based biometric technologies. At present, most gait recognition methods take the whole human body as a unit to establish the spatio-temporal representations. However, we have observed that different parts of human body possess evidently various visual appearances and movement patterns during walking. In the latest literature, employing partial features for human body description has been verified being beneficial to individual recognition. Taken above insights together, we assume that each part of human body needs its own spatio-temporal expression. Then, we propose a novel part-based model GaitPart and get two aspects effect of boosting the performance: On the one hand, Focal Convolution Layer, a new applying of convolution, is presented to enhance the fine-grained learning of the part-level spatial features. On the other hand

, the Micro-motion Capture Module (MCM) is proposed and there are several parall
el MCMs in the GaitPart corresponding to the pre-defined parts of the human body
, respectively. It is worth mentioning that the MCM is a novel way of temporal m
odeling for gait task, which focuses on the short-range temporal features rather
 than the redundant long-range features for cycle gait. Experiments on two of th
e most popular public datasets, CASIA-B and OU-MVLP, richly exemplified that our
 method meets a new state-of-the-art on multiple standard benchmarks. The source
 code will be available on https://github.com/ChaoFan96/GaitPart.
****

Defending and Harnessing the Bit-Flip Based Adversarial Weight Attack
Zhezhi He, Adnan Siraj Rakin, Jingtao Li, Chaitali Chakrabarti, Deliang Fan;
 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti
on (CVPR), 2020, pp. 14095-14103
Recently, a new paradigm of the adversarial attack on the quantized neural netwo
rk weights has attracted great attention, namely, the Bit-Flip based adversarial
 weight attack, aka. Bit-Flip Attack (BFA). BFA has shown extraordinary attackin
g ability, where the adversary can malfunction a quantized Deep Neural Network (
DNN) as a random guess, through malicious bit-flips on a small set of vulnerable
 weight bits (e.g., 13 out of 93 millions bits of 8-bit quantized ResNet-18). Ho
wever, there are no effective defensive methods to enhance the fault-tolerance c
apability of DNN against such BFA. In this work, we conduct comprehensive invest
igations on BFA and propose to leverage binarization-aware training and its rela
xation -- piece-wise clustering as simple and effective countermeasures to BFA.
The experiments show that, for BFA to achieve the identical prediction accuracy
degradation (e.g., below 11% on CIFAR-10), it requires 19.3x and 480.1x more eff
ective malicious bit-flips on ResNet-20 and VGG-11 respectively, compared to def
end-free counterparts.
****

Dense Regression Network for Video Grounding
Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, Chuang Ga
n; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni
tion (CVPR), 2020, pp. 10287-10296
We address the problem of video grounding from natural language queries. The key
 challenge in this task is that one training video might only contain a few anno
tated starting/ending frames that can be used as positive examples for model tra
ining. Most conventional approaches directly train a binary classifier using suc
h imbalance data, thus achieving inferior results. The key idea of this paper is
 to use the distances between the frame within the ground truth and the starting
 (ending) frame as dense supervisions to improve the video grounding accuracy. S
pecifically, we design a novel dense regression network (DRN) to regress the dis
tances from each frame to the starting (ending) frame of the video segment descr
ibed by the query. We also propose a simple but effective IoU regression head mo
dule to explicitly consider the localization quality of the grounding results (i
.e., the IoU between the predicted location and the ground truth). Experimental
results show that our approach significantly outperforms state-of-the-arts on th
ree datasets (i.e., Charades-STA, ActivityNet-Captions, and TACoS).
****

TITAN: Future Forecast Using Action Priors
Srikanth Malla, Behzad Dariush, Chiho Choi; Proceedings of the IEEE/CVF Confer
ence on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11186-11196
We consider the problem of predicting the future trajectory of scene agents from
 egocentric views obtained from a moving platform. This problem is important in
a variety of domains, particularly for autonomous systems making reactive or str
ategic decisions in navigation. In an attempt to address this problem, we introd
uce TITAN (Trajectory Inference using Targeted Action priors Network), a new mod
el that incorporates prior positions, actions, and context to forecast future tr
ajectory of agents and future ego-motion. In the absence of an appropriate datas
et for this task, we created the TITAN dataset that consists of 700 labeled vide
o-clips (with odometry) captured from a moving vehicle on highly interactive urb
an traffic scenes in Tokyo. Our dataset includes 50 labels including vehicle sta

tes and actions, pedestrian age groups, and targeted pedestrian action attributes that are organized hierarchically corresponding to atomic, simple/complex-contextual, transportive, and communicative actions. To evaluate our model, we conducted extensive experiments on the TITAN dataset, revealing significant performance improvement against baselines and state-of-the-art algorithms. We also report promising results from our Agent Importance Mechanism (AIM), a module which provides insight into assessment of perceived risk by calculating the relative influence of each agent on the future ego-trajectory. The dataset is available at https://usa.honda-ri.com/titan

****

Camera On-Boarding for Person Re-Identification Using Hypothesis Transfer Learning

Sk Miraj Ahmed, Aske R. Lejbolle, Rameswar Panda, Amit K. Roy-Chowdhury; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12144-12153

Most of the existing approaches for person re-identification consider a static setting where the number of cameras in the network is fixed. An interesting direction, which has received little attention, is to explore the dynamic nature of a camera network, where one tries to adapt the existing re-identification models after on-boarding new cameras, with little additional effort. There have been a few recent methods proposed in person re-identification that attempt to address this problem by assuming the labeled data in the existing network is still available while adding new cameras. This is a strong assumption since there may exist some privacy issues for which one may not have access to those data. Rather, based on the fact that it is easy to store the learned re-identifications models, which mitigates any data privacy concern, we develop an efficient model adaptation approach using hypothesis transfer learning that aims to transfer the knowledge using only source models and limited labeled data, but without using any source camera data from the existing network. Our approach minimizes the effect of negative transfer by finding an optimal weighted combination of multiple source models for transferring the knowledge. Extensive experiments on four challenging benchmark datasets with variable number of cameras well demonstrate the efficacy of our proposed approach over state-of-the-art methods.

****

EfficientDet: Scalable and Efficient Object Detection

Mingxing Tan, Ruoming Pang, Quoc V. Le; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10781-10790

Model efficiency has become increasingly important in computer vision. In this paper, we systematically study neural network architecture design choices for object detection and propose several key optimizations to improve efficiency. First, we propose a weighted bi-directional feature pyramid network (BiFPN), which allows easy and fast multi-scale feature fusion; Second, we propose a compound scaling method that uniformly scales the resolution, depth, and width for all backbone, feature network, and box/class prediction networks at the same time. Based on these optimizations and EfficientNet backbones, we have developed a new family of object detectors, called EfficientDet, which consistently achieve much better efficiency than prior art across a wide spectrum of resource constraints. In particular, with single-model and single-scale, our EfficientDetD7 achieves state-of-the-art 52.2 AP on COCO test-dev with 52M parameters and 325B FLOPs, being 4x - 9x smaller and using 13x - 42x fewer FLOPs than previous detector.

****

Understanding Adversarial Examples From the Mutual Influence of Images and Perturbations

Chaoning Zhang, Philipp Benz, Tooba Imtiaz, In So Kweon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14521-14530

A wide variety of works have explored the reason for the existence of adversarial examples, but there is no consensus on the explanation. We propose to treat the DNN logits as a vector for feature representation, and exploit them to analyze the mutual influence of two independent inputs based on the Pearson correlation

coefficient (PCC). We utilize this vector representation to understand adversarial examples by disentangling the clean images and adversarial perturbations, and analyze their influence on each other. Our results suggest a new perspective towards the relationship between images and universal perturbations: Universal perturbations contain dominant features, and images behave like noise to them. This feature perspective leads to a new method for generating targeted universal adversarial perturbations using random source images. We are the first to achieve the challenging task of a targeted universal attack without utilizing original training data. Our approach using a proxy dataset achieves comparable performance to the state-of-the-art baselines which utilize the original training dataset.
****

Can Weight Sharing Outperform Random Architecture Search? An Investigation With TuNAS

Gabriel Bender, Hanxiao Liu, Bo Chen, Grace Chu, Shuyang Cheng, Pieter-Jan Kindermans, Quoc V. Le; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14323-14332

Efficient Neural Architecture Search methods based on weight sharing have shown good promise in democratizing Neural Architecture Search for computer vision models. There is, however, an ongoing debate whether these efficient methods are significantly better than random search. Here we perform a thorough comparison between efficient and random search methods on a family of progressively larger and more challenging search spaces for image classification and detection on ImageNet and COCO. While the efficacies of both methods are problem-dependent, our experiments demonstrate that there are large, realistic tasks where efficient search methods can provide substantial gains over random search. In addition, we propose and evaluate techniques which improve the quality of searched architectures and reduce the need for manual hyper-parameter tuning.
****

RMP-SNN: Residual Membrane Potential Neuron for Enabling Deeper High-Accuracy and Low-Latency Spiking Neural Network

Bing Han, Gopalakrishnan Srinivasan, Kaushik Roy; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13558-13567

Spiking Neural Networks (SNNs) have recently attracted significant research interest as the third generation of artificial neural networks that can enable low-power event-driven data analytics. The best performing SNNs for image recognition tasks are obtained by converting a trained Analog Neural Network (ANN), consisting of Rectified Linear Units (ReLU), to SNN composed of integrate-and-fire neurons with "proper" firing thresholds. The converted SNNs typically incur loss in accuracy compared to that provided by the original ANN and require sizable number of inference time-steps to achieve the best accuracy. We find that performance degradation in the converted SNN stems from using "hard reset" spiking neuron that is driven to fixed reset potential once its membrane potential exceeds the firing threshold, leading to information loss during SNN inference. We propose ANN-SNN conversion using "soft reset" spiking neuron model, referred to as Residual Membrane Potential (RMP) spiking neuron, which retains the "residual" membrane potential above threshold at the firing instants. We demonstrate near loss-less ANN-SNN conversion using RMP neurons for VGG-16, ResNet-20, and ResNet-34 SNNs on challenging datasets including CIFAR-10 (93.63% top-1), CIFAR-100 (70.93% top-1), and ImageNet (73.09% top-1 accuracy). Our results also show that RMP-SNN surpasses the best inference accuracy provided by the converted SNN with "hard reset" spiking neurons using 2-8 times fewer inference time-steps across network architectures and datasets.
****

Adversarial Feature Hallucination Networks for Few-Shot Learning

Kai Li, Yulun Zhang, Kunpeng Li, Yun Fu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13470-13479

The recent flourish of deep learning in various tasks is largely accredited to the rich and accessible labeled data. Nonetheless, massive supervision remains a luxury for many real applications, boosting great interest in label-scarce techn

iques such as few-shot learning (FSL), which aims to learn concept of new classes with a few labeled samples. A natural approach to FSL is data augmentation and many recent works have proved the feasibility by proposing various data synthesis models. However, these models fail to well secure the discriminability and diversity of the synthesized data and thus often produce undesirable results. In this paper, we propose Adversarial Feature Hallucination Networks (AFHN) which is based on conditional Wasserstein Generative Adversarial networks (cWGAN) and hallucinates diverse and discriminative features conditioned on the few labeled samples. Two novel regularizers, i.e., the classification regularizer and the anti-collapse regularizer, are incorporated into AFHN to encourage discriminability and diversity of the synthesized features, respectively. Ablation study verifies the effectiveness of the proposed cWGAN based feature hallucination framework and the proposed regularizers. Comparative results on three common benchmark data sets substantiate the superiority of AFHN to existing data augmentation based FSL approaches and other state-of-the-art ones.

****

An Adaptive Neural Network for Unsupervised Mosaic Consistency Analysis in Image Forensics

Quentin Bammey, Rafael Grompone von Gioi, Jean-Michel Morel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14194-14204

Automatically finding suspicious regions in a potentially forged image by splicing, inpainting or copy-move remains a widely open problem. Blind detection neural networks trained on benchmark data are flourishing. Yet, these methods do not provide an explanation of their detections. The more traditional methods try to provide such evidence by pointing out local inconsistencies in the image noise, JPEG compression, chromatic aberration, or in the mosaic. In this paper we develop a blind method that can train directly on unlabelled and potentially forged images to point out local mosaic inconsistencies. To this aim we designed a CNN structure inspired from demosaicing algorithms and directed at classifying image blocks by their position in the image modulo (2 x 2). Creating a diversified benchmark database using varied demosaicing methods, we explore the efficiency of the method and its ability to adapt quickly to any new data.

****

Sign Language Transformers: Joint End-to-End Sign Language Recognition and Translation

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, Richard Bowden; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10023-10033

Prior work on Sign Language Translation has shown that having a mid-level sign gloss representation (effectively recognizing the individual signs) improves the translation performance drastically. In fact, the current state-of-the-art in translation requires gloss level tokenization in order to work. We introduce a novel transformer based architecture that jointly learns Continuous Sign Language Recognition and Translation while being trainable in an end-to-end manner. This is achieved by using a Connectionist Temporal Classification (CTC) loss to bind the recognition and translation problems into a single unified architecture. This joint approach does not require any ground-truth timing information, simultaneously solving two co-dependant sequence-to-sequence learning problems and leads to significant performance gains. We evaluate the recognition and translation performances of our approaches on the challenging RWTH-PHOENIX-Weather-2014T (PHOENIX14T) dataset. We report state-of-the-art sign language recognition and translation results achieved by our Sign Language Transformers. Our translation networks outperform both sign video to spoken language and gloss to spoken language translation models, in some cases more than doubling the performance (9.58 vs. 21.80 BLEU-4 Score). We also share new baseline translation results using transformer networks for several other text-to-text sign language translation tasks.

****

A Context-Aware Loss Function for Action Spotting in Soccer Videos

Anthony Cioppa, Adrien Deliege, Silvio Giancola, Bernard Ghanem, Marc Van Dr

oogenbroeck, Rikke Gade, Thomas B. Moeslund; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13126-13136
In video understanding, action spotting consists in temporally localizing human-induced events annotated with single timestamps. In this paper, we propose a novel loss function that specifically considers the temporal context naturally present around each action, rather than focusing on the single annotated frame to spot. We benchmark our loss on a large dataset of soccer videos, SoccerNet, and achieve an improvement of 12.8% over the baseline. We show the generalization capability of our loss for generic activity proposals and detection on ActivityNet, by spotting the beginning and the end of each activity. Furthermore, we provide an extended ablation study and display challenging cases for action spotting in soccer videos. Finally, we qualitatively illustrate how our loss induces a precise temporal understanding of actions and show how such semantic knowledge can be used for automatic highlights generation.
****

The Edge of Depth: Explicit Constraints Between Segmentation and Depth
Shengjie Zhu, Garrick Brazil, Xiaoming Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13116-13125
In this work we study the mutual benefits of two common computer vision tasks, self-supervised depth estimation and semantic segmentation from images. For example, to help unsupervised monocular depth estimation, constraint from semantic segmentation has been explored implicitly such as sharing and transforming features. In contrast, we propose to explicitly measure the border consistency between segmentation and depth and minimize it in a greedy manner by iteratively supervising the network towards a locally optimal solution. Partially this is motivated by our observation that semantic segmentation even trained with limited ground truth (200 images of KITTI) can offer more accurate border than that of any (monocular or stereo) image-based depth estimation. Through extensive experiments, our proposed approach advance the state of the art on unsupervised monocular depth estimation in the KITTI benchmark.
****

Label Distribution Learning on Auxiliary Label Space Graphs for Facial Expression Recognition
Shikai Chen, Jianfeng Wang, Yuedong Chen, Zhongchao Shi, Xin Geng, Yong Rui; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13984-13993
Many existing studies reveal that annotation inconsistency widely exists among a variety of facial expression recognition (FER) datasets. The reason might be the subjectivity of human annotators and the ambiguous nature of the expression labels. One promising strategy tackling such a problem is a recently proposed learning paradigm called Label Distribution Learning (LDL), which allows multiple labels with different intensity to be linked to one expression. However, it is often impractical to directly apply label distribution learning because numerous existing datasets only contain one-hot labels rather than label distributions. To solve the problem, we propose a novel approach named Label Distribution Learning on Auxiliary Label Space Graphs(LDL-ALSG) that leverages the topological information of the labels from related but more distinct tasks, such as action unit recognition and facial landmark detection. The underlying assumption is that facial images should have similar expression distributions to their neighbours in the label space of action unit recognition and facial landmark detection. Our proposed method is evaluated on a variety of datasets and outperforms those state-of-the-art methods consistently with a huge margin.
****

Cross-Modality Person Re-Identification With Shared-Specific Feature Transfer
Yan Lu, Yue Wu, Bin Liu, Tianzhu Zhang, Baopu Li, Qi Chu, Nenghai Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13379-13389
Cross-modality person re-identification (cm-ReID) is a challenging but key technology for intelligent video analysis. Existing works mainly focus on learning modality-shared representation by embedding different modalities into a same featu

re space, lowering the upper bound of feature distinctiveness. In this paper, we tackle the above limitation by proposing a novel cross-modality shared-specific feature transfer algorithm (termed cm-SSFT) to explore the potential of both the modality-shared information and the modality-specific characteristics to boost the reidentification performance. We model the affinities of different modality samples according to the shared features and then transfer both shared and specific features among and across modalities. We also propose a complementary feature learning strategy including modality adaption, project adversarial learning and reconstruction enhancement to learn discriminative and complementary shared and specific features of each modality, respectively. The entire cmSSFTalgorithm can be trained in an end-to-end manner. We conducted comprehensive experiments to validate the superiority ofthe overall algorithm and the effectiveness ofeach component. The proposed algorithm significantly outperforms state-of-the-arts by 22.5% and 19.3% mAP on the two mainstream benchmark datasets SYSU-MM01 and RegDB, respectively.

****

Learning a Unified Sample Weighting Network for Object Detection

Qi Cai, Yingwei Pan, Yu Wang, Jingen Liu, Ting Yao, Tao Mei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14173-14182

Region sampling or weighting is significantly important to the success of modern region-based object detectors. Unlike some previous works, which only focus on "hard" samples when optimizing the objective function, we argue that sample weighting should be data-dependent and task-dependent. The importance of a sample for the objective function optimization is determined by its uncertainties to both object classification and bounding box regression tasks. To this end, we devise a general loss function to cover most region-based object detectors with various sampling strategies, and then based on it we propose a unified sample weighting network to predict a sample's task weights. Our framework is simple yet effective. It leverages the samples' uncertainty distributions on classification loss, regression loss, IoU, and probability score, to predict sample weights. Our approach has several advantages: (i). It jointly learns sample weights for both classification and regression tasks, which differentiates it from most previous work. (ii). It is a data-driven process, so it avoids some manual parameter tuning. (iii). It can be effortlessly plugged into most object detectors and achieves noticeable performance improvements without affecting their inference time. Our approach has been thoroughly evaluated with recent object detection frameworks and it can consistently boost the detection accuracy. Code has been made available at https://github.com/caiqi/sample-weighting-network.

****

Joint Semantic Segmentation and Boundary Detection Using Iterative Pyramid Contexts

Mingmin Zhen, Jinglu Wang, Lei Zhou, Shiwei Li, Tianwei Shen, Jiaxiang Shang, Tian Fang, Long Quan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13666-13675

In this paper, we present a joint multi-task learning framework for semantic segmentation and boundary detection. The critical component in the framework is the iterative pyramid context module (PCM), which couples two tasks and stores the shared latent semantics to interact between the two tasks. For semantic boundary detection, we propose the novel spatial gradient fusion to suppress non-semantic edges. As semantic boundary detection is the dual task of semantic segmentation, we introduce a loss function with boundary consistency constraint to improve the boundary pixel accuracy for semantic segmentation. Our extensive experiments demonstrate superior performance over state-of-the-art works, not only in semantic segmentation but also in semantic boundary detection. In particular, a mean IoU score of 81.8% on Cityscapes test set is achieved without using coarse data or any external data for semantic segmentation. For semantic boundary detection, we improve over previous state-of-the-art works by 9.9% in terms of AP and 6.8% in terms of MF(ODS).

****

SLV: Spatial Likelihood Voting for Weakly Supervised Object Detection
Ze Chen, Zhihang Fu, Rongxin Jiang, Yaowu Chen, Xian-Sheng Hua; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12995-13004
Based on the framework of multiple instance learning (MIL), tremendous works have promoted the advances of weakly supervised object detection (WSOD). However, most MIL-based methods tend to localize instances to their discriminative parts instead of the whole content. In this paper, we propose a spatial likelihood voting (SLV) module to converge the proposal localizing process without any bounding box annotations. Specifically, all region proposals in a given image play the role of voters every iteration during training, voting for the likelihood of each category in spatial dimensions. After dilating alignment on the area with large likelihood values, the voting results are regularized as bounding boxes, being used for the final classification and localization. Based on SLV, we further propose an end-to-end training framework for multi-task learning. The classification and localization tasks promote each other, which further improves the detection performance. Extensive experiments on the PASCAL VOC 2007 and 2012 datasets demonstrate the superior performance of SLV.
****
Robust Superpixel-Guided Attentional Adversarial Attack
Xiaoyi Dong, Jiangfan Han, Dongdong Chen, Jiayang Liu, Huanyu Bian, Zehua Ma, Hongsheng Li, Xiaogang Wang, Weiming Zhang, Nenghai Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12895-12904
Deep Neural Networks are vulnerable to adversarial samples, which can fool classifiers by adding small perturbations onto the original image. Since the pioneering optimization-based adversarial attack method, many following methods have been proposed in the past several years. However most of these methods add perturbations in a "pixel-wise" and "global" way. Firstly, because of the contradiction between the local smoothness of natural images and the noisy property of these adversarial perturbations, this "pixel-wise" way makes these methods not robust to image processing based defense methods and steganalysis based detection methods. Secondly, we find adding perturbations to the background is less useful than to the salient object, thus the "global" way is also not optimal. Based on these two considerations, we propose the first robust superpixel-guided attentional adversarial attack method. Specifically, the adversarial perturbations are only added to the salient regions and guaranteed to be same within each superpixel. Through extensive experiments, we demonstrate our method can preserve the attack ability even in this highly constrained modification space. More importantly, compared to existing methods, it is significantly more robust to image processing based defense and steganalysis based detection.
****
MMTM: Multimodal Transfer Module for CNN Fusion
Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L. Iuzzolino, Kazuhito Koishida; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13289-13299
In late fusion, each modality is processed in a separate unimodal Convolutional Neural Network (CNN) stream and the scores of each modality are fused at the end. Due to its simplicity, late fusion is still the predominant approach in many state-of-the-art multimodal applications. In this paper, we present a simple neural network module for leveraging the knowledge from multiple modalities in convolutional neural networks. The proposed unit, named Multimodal Transfer Module (MMTM), can be added at different levels of the feature hierarchy, enabling slow modality fusion. Using squeeze and excitation operations, MMTM utilizes the knowledge of multiple modalities to recalibrate the channel-wise features in each CNN stream. Unlike other intermediate fusion methods, the proposed module could be used for feature modality fusion in convolution layers with different spatial dimensions. Another advantage of the proposed method is that it could be added among unimodal branches with minimum changes in the their network architectures, allowing each branch to be initialized with existing pretrained weights. Experimen

tal results show that our framework improves the recognition accuracy of well-kn
own multimodal networks. We demonstrate state-of-the-art or competitive performa
nce on four datasets that span the task domains of dynamic hand gesture recognit
ion, speech enhancement, and action recognition with RGB and body joints.
****

Optical Flow in Dense Foggy Scenes Using Semi-Supervised Learning
Wending Yan, Aashish Sharma, Robby T. Tan; Proceedings of the IEEE/CVF Confere
nce on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13259-13268
In dense foggy scenes, existing optical flow methods are erroneous. This is due
to the degradation caused by dense fog particles that break the optical flow bas
ic assumptions such as brightness and gradient constancy. To address the problem
, we introduce a semi-supervised deep learning technique that employs real fog i
mages without optical flow ground-truths in the training process. Our network in
tegrates the domain transformation and optical flow networks in one framework. I
nitially, given a pair of synthetic fog images, its corresponding clean images a
nd optical flow ground-truths, in one training batch we train our network in a s
upervised manner. Subsequently, given a pair of real fog images and a pair of cl
ean images that are not corresponding to each other (unpaired), in the next trai
ning batch, we train our network in an unsupervised manner. We then alternate th
e training of synthetic and real data iteratively. We use real data without grou
nd-truths, since to have ground-truths in such conditions is intractable, and al
so to avoid the overfitting problem of synthetic data training, where the knowle
dge learned on synthetic data cannot be generalized to real data testing. Togeth
er with the network architecture design, we propose a new training strategy that
combines supervised synthetic-data training and unsupervised real-data training
. Experimental results show that our method is effective and outperforms the sta
te-of-the-art methods in estimating optical flow in dense foggy scenes.
****

Learning Memory-Guided Normality for Anomaly Detection
Hyunjong Park, Jongyoun Noh, Bumsub Ham; Proceedings of the IEEE/CVF Conferenc
e on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14372-14381
We address the problem of anomaly detection, that is, detecting anomalous events
in a video sequence. Anomaly detection methods based on convolutional neural ne
tworks (CNNs) typically leverage proxy tasks, such as reconstructing input video
frames, to learn models describing normality without seeing anomalous samples a
t training time, and quantify the extent of abnormalities using the reconstructi
on error at test time. The main drawbacks of these approaches are that they do n
ot consider the diversity of normal patterns explicitly, and the powerful repres
entation capacity of CNNs allows to reconstruct abnormal video frames. To addres
s this problem, we present an unsupervised learning approach to anomaly detectio
n that considers the diversity of normal patterns explicitly, while lessening th
e representation capacity of CNNs. To this end, we propose to use a memory modul
e with a new update scheme where items in the memory record prototypical pattern
s of normal data. We also present novel feature compactness and separateness los
ses to train the memory, boosting the discriminative power of both memory items
and deeply learned features from normal data. Experimental results on standard b
enchmarks demonstrate the effectiveness and efficiency of our approach, which ou
tperforms the state of the art.
****

MLCVNet: Multi-Level Context VoteNet for 3D Object Detection
Qian Xie, Yu-Kun Lai, Jing Wu, Zhoutao Wang, Yiming Zhang, Kai Xu, Jun Wan
g; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni
tion (CVPR), 2020, pp. 10447-10456
In this paper, we address the 3D object detection task by capturing multi-level
contextual information with the self-attention mechanism and multi-scale feature
fusion. Most existing 3D object detection methods recognize objects individuall
y, without giving any consideration on contextual information between these obje
cts. Comparatively, we propose Multi-Level Context VoteNet (MLCVNet) to recogniz
e 3D objects correlatively, building on the state-of-the-art VoteNet. We introdu
ce three context modules into the voting and classifying stages of VoteNet to en

code contextual information at different levels. Specifically, a Patch-to-Patch Context (PPC) module is employed to capture contextual information between the p oint patches, before voting for their corresponding object centroid points. Subs equently, an Object-to-Object Context (OOC) module is incorporated before the pr oposal and classification stage, to capture the contextual information between o bject candidates. Finally, a Global Scene Context (GSC) module is designed to le arn the global scene context. We demonstrate these by capturing contextual infor mation at patch, object and scene levels. Our method is an effective way to prom ote detection accuracy, achieving new state-of-the-art detection performance on challenging 3D object detection datasets, i.e., SUN RGBD and ScanNet. We also re lease our code at https://github.com/NUAAXQ/MLCVNet.
****

SQuINTing at VQA Models: Introspecting VQA Models With Sub-Questions
Ramprasaath R. Selvaraju, Purva Tendulkar, Devi Parikh, Eric Horvitz, Marco Tulio Ribeiro, Besmira Nushi, Ece Kamar; Proceedings of the IEEE/CVF Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10003-10011
Existing VQA datasets contain questions with varying levels of complexity. While the majority of questions in these datasets require perception for recognizing existence, properties, and spatial relationships of entities, a significant port ion of questions pose challenges that correspond to reasoning tasks - tasks that can only be answered through a synthesis of perception and knowledge about the world, logic and / or reasoning. Analyzing performance across this distinction a llows us to notice when existing VQA models have consistency issues - they answe r the reasoning questions correctly but fail on associated low-level perception questions. For example, in Figure 1, models answer the complex reasoning questio n "Is the banana ripe enough to eat?" correctly, but fail on the associated perc eption question "Are the bananas mostly green or yellow?" indicating that the mo del likely answered the reasoning question correctly but for the wrong reason. W e quantify the extent to which this phenomenon occurs by creating a new Reasonin g split of the VQA dataset and collecting VQAintrospect, a new dataset1 which cu rrently consists of 200K new perception questions which serve as sub questions c orresponding to the set of perceptual tasks needed to effectively answer the com plex reasoning questions in the Reasoning split. Our evaluation shows that state -of-the-art VQA models have comparable performance in answering perception and r easoning questions, but suffer from consistency problems. To address this shortc oming, we propose an approach called Sub-Question Importance-aware Network Tunin g (SQuINT), which encourages the model to attend to the same parts of the image when answering the reasoning question and the perception sub question. We show t hat SQuINT improves model consistency by 7%, also marginally improving performa nce on the Reasoning questions in VQA, while also displaying better attention ma ps.
****

VectorNet: Encoding HD Maps and Agent Dynamics From Vectorized Representation
Jiyang Gao, Chen Sun, Hang Zhao, Yi Shen, Dragomir Anguelov, Congcong Li, Cordelia Schmid; Proceedings of the IEEE/CVF Conference on Computer Vision and P attern Recognition (CVPR), 2020, pp. 11525-11533
Behavior prediction in dynamic, multi-agent systems is an important problem in t he context of self-driving cars, due to the complex representations and interact ions of road components, including moving agents (e.g. pedestrians and vehicles) and road context information (e.g. lanes, traffic lights). This paper introduce s VectorNet, a hierarchical graph neural network that first exploits the spatial locality of individual road components represented by vectors and then models t he high-order interactions among all components. In contrast to most recent appr oaches, which render trajectories of moving agents and road context information as bird-eye images and encode them with convolutional neural networks (ConvNets) , our approach operates on the primitive vector representation. By operating on the vectorized high definition (HD) maps and agent trajectories, we avoid lossy rendering and computationally intensive ConvNet encoding steps. To further boost VectorNet's capability in learning context features, we propose a novel auxilia ry task to recover the randomly masked out map entities and agent trajectories b

ased on their context. We evaluate VectorNet on our in-house behavior prediction benchmark and the recently released Argoverse forecasting dataset. Our method achieves on par or better performance than the competitive rendering approach on both benchmarks while saving over 70% of the model parameters with an order of magnitude reduction in FLOPs. It also obtains state-of-the-art performance on the Argoverse dataset.
****

Through Fog High-Resolution Imaging Using Millimeter Wave Radar
Junfeng Guan, Sohrab Madani, Suraj Jog, Saurabh Gupta, Haitham Hassanieh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11464-11473
This paper demonstrates high-resolution imaging using millimeter Wave (mmWave) radars that can function even in dense fog. We leverage the fact that mmWave signals have favorable propagation characteristics in low visibility conditions, unlike optical sensors like cameras and LiDARs which cannot penetrate through dense fog. Millimeter-wave radars, however, suffer from very low resolution, specularity, and noise artifacts. We introduce HawkEye, a system that leverages a cGAN architecture to recover high-frequency shapes from raw low-resolution mmWave heat-maps. We propose a novel design that addresses challenges specific to the structure and nature of the radar signals involved. We also develop a data synthesizer to aid with large-scale dataset generation for training. We implement our system on a custom-built mmWave radar platform and demonstrate performance improvement over both standard mmWave radars and other competitive baselines.
****

Self-Supervised Learning of Video-Induced Visual Invariances
Michael Tschannen, Josip Djolonga, Marvin Ritter, Aravindh Mahendran, Neil Houlsby, Sylvain Gelly, Mario Lucic; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13806-13815
We propose a general framework for self-supervised learning of transferable visual representations based on Video-Induced Visual Invariances (VIVI). We consider the implicit hierarchy present in the videos and make use of (i) frame-level invariances (e.g. stability to color and contrast perturbations), (ii) shot/clip-level invariances (e.g. robustness to changes in object orientation and lighting conditions), and (iii) video-level invariances (semantic relationships of scenes across shots/clips), to define a holistic self-supervised loss. Training models using different variants of the proposed framework on videos from the YouTube-8M (YT8M) data set, we obtain state-of-the-art self-supervised transfer learning results on the 19 diverse downstream tasks of the Visual Task Adaptation Benchmark (VTAB), using only 1000 labels per task. We then show how to co-train our models jointly with labeled images, outperforming an ImageNet-pretrained ResNet-50 by 0.8 points with 10x fewer labeled images, as well as the previous best supervised model by 3.7 points using the full ImageNet data set.
****

Butterfly Transform: An Efficient FFT Based Neural Architecture Design
Keivan Alizadeh vahid, Anish Prabhu, Ali Farhadi, Mohammad Rastegari; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12024-12033
In this paper, we show that extending the butterfly operations from the FFT algorithm to a general Butterfly Transform (BFT) can be beneficial in building an efficient block structure for CNN designs. Pointwise convolutions, which we refer to as channel fusions, are the main computational bottleneck in the state-of-the-art efficient CNNs (e.g. MobileNets). We introduce a set of criterion for channel fusion, and prove that BFT yields an asymptotically optimal FLOP count with respect to these criteria. By replacing pointwise convolutions with BFT, we reduce the computational complexity of these layers from O(n^2) to O(n log n) with respect to the number of channels. Our experimental evaluations show that our method results in significant accuracy gains across a wide range of network architectures, especially at low FLOP ranges. For example, BFT results in up to a 6.75% absolute Top-1 improvement for MobileNetV1, 4.4 % for ShuffleNet V2 and 5.4% for MobileNetV3 on ImageNet under a similar number of FLOPS. Notably, ShuffleNet-V2

+BFT outperforms state-of-the-art architecture search methods MNasNet, FBNet and MobilenetV3 in the low FLOP regime.
****

Cross-Domain Detection via Graph-Induced Prototype Alignment
Minghao Xu, Hang Wang, Bingbing Ni, Qi Tian, Wenjun Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12355-12364
Applying the knowledge of an object detector trained on a specific domain directly onto a new domain is risky, as the gap between two domains can severely degrade model's performance. Furthermore, since different instances commonly embody distinct modal information in object detection scenario, the feature alignment of source and target domain is hard to be realized. To mitigate these problems, we propose a Graph-induced Prototype Alignment (GPA) framework to seek for category-level domain alignment via elaborate prototype representations. In the nutshell, more precise instance-level features are obtained through graph-based information propagation among region proposals, and, on such basis, the prototype representation of each class is derived for category-level domain alignment. In addition, in order to alleviate the negative effect of class-imbalance on domain adaptation, we design a Class-reweighted Contrastive Loss to harmonize the adaptation training process. Combining with Faster R-CNN, the proposed framework conducts feature alignment in a two-stage manner. Comprehensive results on various cross-domain detection tasks demonstrate that our approach outperforms existing methods with a remarkable margin. Our code is available at https://github.com/ChrisAllenMing/GPA-detection.
****

What Makes Training Multi-Modal Classification Networks Hard?
Weiyao Wang, Du Tran, Matt Feiszli; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12695-12705
Consider end-to-end training of a multi-modal vs. a uni-modal network on a task with multiple input modalities: the multi-modal network receives more information, so it should match or outperform its uni-modal counterpart. In our experiments, however, we observe the opposite: the best uni-modal network can outperform the multi-modal network. This observation is consistent across different combinations of modalities and on different tasks and benchmarks for video classifications. This paper identifies two main causes for this performance drop: first, multi-modal networks are often prone to overfitting due to increased capacity. Second, different modalities overfit and generalize at different rates, so training them jointly with a single optimization strategy is sub-optimal. We address these two problems with a technique we call Gradient-Blending, which computes an optimal blending of modalities based on their overfitting behaviors. We demonstrate that Gradient Blending outperforms widely-used baselines for avoiding overfitting and achieves state-of-the-art accuracy on various tasks including human action recognition, ego-centric action recognition, and acoustic event detection.
****

Sparse Layered Graphs for Multi-Object Segmentation
Niels Jeppesen, Anders N. Christensen, Vedrana A. Dahl, Anders B. Dahl; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12777-12785
We introduce the novel concept of a Sparse Layered Graph (SLG) for s-t graph cut segmentation of image data. The concept is based on the widely used Ishikawa layered technique for multi-object segmentation, which allows explicit object interactions, such as containment and exclusion with margins. However, the spatial complexity of the Ishikawa technique limits its use for many segmentation problems. To solve this issue, we formulate a general method for adding containment and exclusion interaction constraints to layered graphs. Given some prior knowledge, we can create a SLG, which is often orders of magnitude smaller than traditional Ishikawa graphs, with identical segmentation results. This allows us to solve many problems that could previously not be solved using general graph cut algorithms. We then propose three algorithms for further reducing the spatial complexity of SLGs, by using ordered multi-column graphs. In our experiments, we show t

hat SLGs, and in particular ordered multi-column SLGs, can produce high-quality segmentation results using extremely simple data terms. We also show the scalability of ordered multi-column SLGs, by segmenting a high-resolution volume with several hundred interacting objects.
****

Few-Shot Class-Incremental Learning
Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, Yihong Gong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12183-12192
The ability to incrementally learn new classes is crucial to the development of real-world artificial intelligence systems. In this paper, we focus on a challenging but practical few-shot class-incremental learning (FSCIL) problem. FSCIL requires CNN models to incrementally learn new classes from very few labelled samples, without forgetting the previously learned ones. To address this problem, we represent the knowledge using a neural gas (NG) network, which can learn and preserve the topology of the feature manifold formed by different classes. On this basis, we propose the TOpology-Preserving knowledge InCrementer (TOPIC) framework. TOPIC mitigates the forgetting of the old classes by stabilizing NG's topology and improves the representation learning for few-shot new classes by growing and adapting NG to new training samples. Comprehensive experimental results demonstrate that our proposed method significantly outperforms other state-of-the-art class-incremental learning methods on CIFAR100, miniImageNet, and CUB200 datasets.
****

Exploring Bottom-Up and Top-Down Cues With Attentive Learning for Webly Supervised Object Detection
Zhonghua Wu, Qingyi Tao, Guosheng Lin, Jianfei Cai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12936-12945
Fully supervised object detection has achieved great success in recent years. However, abundant bounding boxes annotations are needed for training a detector for novel classes. To reduce the human labeling effort, we propose a novel webly supervised object detection (WebSOD) method for novel classes which only requires the web images without further annotations. Our proposed method combines bottom-up and top-down cues for novel class detection. Within our approach, we introduce a bottom-up mechanism based on the well-trained fully supervised object detector (i.e. Faster RCNN) as an object region estimator for web images by recognizing the common objectiveness shared by base and novel classes. With the estimated regions on the web images, we then utilize the top-down attention cues as the guidance for region classification. Furthermore, we propose a residual feature refinement (RFR) block to tackle the domain mismatch between web domain and the target domain. We demonstrate our proposed method on PASCAL VOC dataset with three different novel/base splits. Without any target-domain novel-class images and annotations, our proposed webly supervised object detection model is able to achieve promising performance for novel classes. Moreover, we also conduct transfer learning experiments on large scale ILSVRC 2013 detection dataset and achieve state-of-the-art performance.
****

SpineNet: Learning Scale-Permuted Backbone for Recognition and Localization
Xianzhi Du, Tsung-Yi Lin, Pengchong Jin, Golnaz Ghiasi, Mingxing Tan, Yin Cui, Quoc V. Le, Xiaodan Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11592-11601
Convolutional neural networks typically encode an input image into a series of intermediate features with decreasing resolutions. While this structure is suited to classification tasks, it does not perform well for tasks requiring simultaneous recognition and localization (e.g., object detection). The encoder-decoder architectures are proposed to resolve this by applying a decoder network onto a backbone model designed for classification tasks. In this paper, we argue encoder-decoder architecture is ineffective in generating strong multi-scale features because of the scale-decreased backbone. We propose SpineNet, a backbone with sca

le-permuted intermediate features and cross-scale connections that is learned on an object detection task by Neural Architecture Search. Using similar building blocks, SpineNet models outperform ResNet-FPN models by 3%+ AP at various scales while using 10-20% fewer FLOPs. In particular, SpineNet-190 achieves 52.1% AP on COCO, attaining the new state-of-the-art performance for single model object detection without test-time augmentation. SpineNet can transfer to classification tasks, achieving 5% top-1 accuracy improvement on a challenging iNaturalist fine-grained dataset. Code is at: https://github.com/tensorflow/tpu/tree/master/models/official/detection.
****

LatentFusion: End-to-End Differentiable Reconstruction and Rendering for Unseen Object Pose Estimation
Keunhong Park, Arsalan Mousavian, Yu Xiang, Dieter Fox; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10710-10719
Current 6D object pose estimation methods usually require a 3D model for each object. These methods also require additional training in order to incorporate new objects. As a result, they are difficult to scale to a large number of objects and cannot be directly applied to unseen objects. We propose a novel framework for 6D pose estimation of unseen objects. We present a network that reconstructs a latent 3D representation of an object using a small number of reference views at inference time. Our network is able to render the latent 3D representation from arbitrary views. Using this neural renderer, we directly optimize for pose given an input image. By training our network with a large number of 3D shapes for reconstruction and rendering, our network generalizes well to unseen objects. We present a new dataset for unseen object pose estimation--MOPED. We evaluate the performance of our method for unseen object pose estimation on MOPED as well as the ModelNet and LINEMOD datasets. Our method performs competitively to supervised methods that are trained on those objects. Code and data will be available at https://keunhong.com/publications/latentfusion/
****

Offset Bin Classification Network for Accurate Object Detection
Heqian Qiu, Hongliang Li, Qingbo Wu, Hengcan Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13188-13197
Object detection combines object classification and object localization problems. Most existing object detection methods usually locate objects by leveraging regression networks trained with Smooth L1 loss function to predict offsets between candidate boxes and objects. However, this loss function applies the same penalties on different samples with large errors, which results in suboptimal regression networks and inaccurate offsets. In this paper, we propose an offset bin classification network optimized with cross entropy loss to predict more accurate offsets. It not only provides different penalties for different samples but also avoids the gradient explosion problem caused by the samples with large errors. Specifically, we discretize the continuous offset into a number of bins, and predict the probability of each offset bin. Furthermore, we propose an expectation-based offset prediction and a hierarchical focusing method to improve the prediction precision. Extensive experiments on the PASCAL VOC and MS-COCO datasets demonstrate the effectiveness of our proposed method. Our method outperforms the baseline methods by a large margin.
****

Generating Accurate Pseudo-Labels in Semi-Supervised Learning and Avoiding Overconfident Predictions via Hermite Polynomial Activations
Vishnu Suresh Lokhande, Songwong Tasneeyapant, Abhay Venkatesh, Sathya N. Ravi, Vikas Singh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11435-11443
Rectified Linear Units (ReLUs) are among the most widely used activation function in a broad variety of tasks in vision. Recent theoretical results suggest that despite their excellent practical performance, in various cases, a substitution with basis expansions (e.g., polynomials) can yield significant benefits from b

oth the optimization and generalization perspective. Unfortunately, the existing results remain limited to networks with a couple of layers, and the practical viability of these results is not yet known. Motivated by some of these results, we explore the use of Hermite polynomial expansions as a substitute for ReLUs in deep networks. While our experiments with supervised learning do not provide a clear verdict, we find that this strategy offers considerable benefits in semi-supervised learning (SSL) / transductive learning settings. We carefully develop this idea and show how the use of Hermite polynomials based activations can yield improvements in pseudo-label accuracies and sizable financial savings (due to concurrent runtime benefits). Further, we show via theoretical analysis, that the networks (with Hermite activations) offer robustness to noise and other attractive mathematical properties.

****

MiLeNAS: Efficient Neural Architecture Search via Mixed-Level Reformulation
Chaoyang He, Haishan Ye, Li Shen, Tong Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11993-12002
Many recently proposed methods for Neural Architecture Search (NAS) can be formulated as bilevel optimization. For efficient implementation, its solution requires approximations of second-order methods. In this paper, we demonstrate that gradient errors caused by such approximations lead to suboptimality, in the sense that the optimization procedure fails to converge to a (locally) optimal solution. To remedy this, this paper proposes MiLeNAS, a mixed-level reformulation for NAS that can be optimized efficiently and reliably. It is shown that even when using a simple first-order method on the mixed-level formulation, MiLeNAS can achieve a lower validation error for NAS problems. Consequently, architectures obtained by our method achieve consistently higher accuracies than those obtained from bilevel optimization. Moreover, MiLeNAS proposes a framework beyond DARTS. It is upgraded via model size-based search and early stopping strategies to complete the search process in around 5 hours. Extensive experiments within the convolutional architecture search space validate the effectiveness of our approach.

****

G-TAD: Sub-Graph Localization for Temporal Action Detection
Mengmeng Xu, Chen Zhao, David S. Rojas, Ali Thabet, Bernard Ghanem; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10156-10165
Temporal action detection is a fundamental yet challenging task in video understanding. Video context is a critical cue to effectively detect actions, but current works mainly focus on temporal context, while neglecting semantic context as well as other important context properties. In this work, we propose a graph convolutional network (GCN) model to adaptively incorporate multi-level semantic context into video features and cast temporal action detection as a sub-graph localization problem. Specifically, we formulate video snippets as graph nodes, snippet-snippet correlations as edges, and actions associated with context as target sub-graphs. With graph convolution as the basic operation, we design a GCN block called GCNeXt, which learns the features of each node by aggregating its context and dynamically updates the edges in the graph. To localize each sub-graph, we also design an SGAlign layer to embed each sub-graph into the Euclidean space. Extensive experiments show that G-TAD is capable of finding effective video context without extra supervision and achieves state-of-the-art performance on two detection benchmarks. On ActivityNet-1.3 it obtains an average mAP of 34.09%; on THUMOS14 it reaches 51.6% at IoU@0.5 when combined with a proposal processing method. The code has been made available at https://github.com/frostinassiky/gtad.

****

Learning Saliency Propagation for Semi-Supervised Instance Segmentation
Yanzhao Zhou, Xin Wang, Jianbin Jiao, Trevor Darrell, Fisher Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10307-10316
Instance segmentation is a challenging task for both modeling and annotation. Due to the high annotation cost, modeling becomes more difficult because of the li

mited amount of supervision. We aim to improve the accuracy of the existing inst ance segmentation models by utilizing a large amount of detection supervision. W e propose ShapeProp, which learns to activate the salient regions within the obj ect detection and propagate the areas to the whole instance through an iterative learnable message passing module. ShapeProp can benefit from more bounding box supervision to locate the instances more accurately and utilize the feature acti vations from the larger number of instances to achieve more accurate segmentatio n. We extensively evaluate ShapeProp on three datasets (MS COCO, PASCAL VOC, and BDD100k) with different supervision setups based on both two-stage (Mask R-CNN) and single-stage (RetinaMask) models. The results show our method establishes n ew states of the art for semi-supervised instance segmentation.
****

Dataless Model Selection With the Deep Frame Potential
Calvin Murdock, Simon Lucey; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11257-11265
Choosing a deep neural network architecture is a fundamental problem in applicat ions that require balancing performance and parameter efficiency. Standard appro aches rely on ad-hoc engineering or computationally expensive validation on a sp ecific dataset. We instead attempt to quantify networks by their intrinsic capac ity for unique and robust representations, enabling efficient architecture compa risons without requiring any data. Building upon theoretical connections between deep learning and sparse approximation, we propose the deep frame potential: a measure of coherence that is approximately related to representation stability b ut has minimizers that depend only on network structure. This provides a framewo rk for jointly quantifying the contributions of architectural hyper-parameters s uch as depth, width, and skip connections. We validate its use as a criterion fo r model selection and demonstrate correlation with generalization error on a var iety of common residual and densely connected network architectures.
****

MUXConv: Information Multiplexing in Convolutional Neural Networks
Zhichao Lu, Kalyanmoy Deb, Vishnu Naresh Boddeti; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12044-12 053
Convolutional neural networks have witnessed remarkable improvements in computat ional efficiency in recent years. A key driving force has been the idea of tradi ng-off model expressivity and efficiency through a combination of 1x1 and depth- wise separable convolutions in lieu of a standard convolutional layer. The price of the efficiency, however, is the sub-optimal flow of information across space and channels in the network. To overcome this limitation, we present MUXConv, a layer that is designed to increase the flow of information by progressively mul tiplexing channel and spatial information in the network, while mitigating compu tational complexity. Furthermore, to demonstrate the effectiveness of MUXConv, w e integrate it within an efficient multi-objective evolutionary algorithm to sea rch for the optimal model hyper-parameters while simultaneously optimizing accur acy, compactness, and computational efficiency. On ImageNet, the resulting model s, dubbed MUXNets, match the performance (75.3% top-1 accuracy) and multiply-add operations (218M) of MobileNetV3 while being 1.6x more compact, and outperform other mobile models in all the three criteria. MUXNet also performs well under t ransfer learning and when adapted to object detection. On the ChestX-Ray 14 benc hmark, its accuracy is comparable to the state-of-the-art while being 3.3x more compact and 14x more efficient. Similarly, detection on PASCAL VOC 2007 is 1.2% more accurate, 28% faster and 6% more compact compared to MobileNetV2.
****

Learning to Segment 3D Point Clouds in 2D Image Space
Yecheng Lyu, Xinming Huang, Ziming Zhang; Proceedings of the IEEE/CVF Conferen ce on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12255-12264
In contrast to the literature where local patterns in 3D point clouds are captur ed by customized convolutional operators, in this paper we study the problem of how to effectively and efficiently project such point clouds into a 2D image spa ce so that traditional 2D convolutional neural networks (CNNs) such as U-Net can

be applied for segmentation. To this end, we are motivated by graph drawing and reformulate it as an integer programming problem to learn the topology-preserving graph-to-grid mapping for each individual point cloud. To accelerate the computation in practice, we further propose a novel hierarchical approximate algorithm. With the help of the Delaunay triangulation for graph construction from point clouds and a multi-scale U-Net for segmentation, we manage to demonstrate the state-of-the-art performance on ShapeNet and PartNet, respectively, with significant improvement over the literature. Code is available at https://github.com/Zhang-VISLab.

****

Interactive Image Segmentation With First Click Attention

Zheng Lin, Zhao Zhang, Lin-Zhuo Chen, Ming-Ming Cheng, Shao-Ping Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13339-13348

In the task of interactive image segmentation, users initially click one point to segment the main body of the target object and then provide more points on mislabeled regions iteratively for a precise segmentation. Existing methods treat all interaction points indiscriminately, ignoring the difference between the first click and the remaining ones. In this paper, we demonstrate the critical role of the first click about providing the location and main body information of the target object. A deep framework, named First Click Attention Network (FCA-Net), is proposed to make better use of the first click. In this network, the interactive segmentation result can be much improved with the following benefits: focus invariance, location guidance, and error-tolerant ability. We then put forward a click-based loss function and a structural integrity strategy for better segmentation effect. The visualized segmentation results and sufficient experiments on five datasets demonstrate the importance of the first click and the superiority of our FCA-Net.

****

Attention Convolutional Binary Neural Tree for Fine-Grained Visual Categorization

Ruyi Ji, Longyin Wen, Libo Zhang, Dawei Du, Yanjun Wu, Chen Zhao, Xianglong Liu, Feiyue Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10468-10477

Fine-grained visual categorization (FGVC) is an important but challenging task due to high intra-class variances and low inter-class variances caused by deformation, occlusion, illumination, etc. An attention convolutional binary neural tree architecture is presented to address those problems for weakly supervised FGVC. Specifically, we incorporate convolutional operations along edges of the tree structure, and use the routing functions in each node to determine the root-to-leaf computational paths within the tree. The final decision is computed as the summation of the predictions from leaf nodes. The deep convolutional operations learn to capture the representations of objects, and the tree structure characterizes the coarse-to-fine hierarchical feature learning process. In addition, we use the attention transformer module to enforce the network to capture discriminative features. The negative log-likelihood loss is used to train the entire network in an end-to-end fashion by SGD with back-propagation. Several experiments on the CUB-200-2011, Stanford Cars and Aircraft datasets demonstrate that the proposed method performs favorably against the state-of-the-arts.

****

Dynamic Convolution: Attention Over Convolution Kernels

Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, Zicheng Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11030-11039

Light-weight convolutional neural networks (CNNs) suffer performance degradation as their low computational budgets constrain both the depth (number of convolution layers) and the width (number of channels) of CNNs, resulting in limited representation capability. To address this issue, we present Dynamic Convolution, a new design that increases model complexity without increasing the network depth or width. Instead of using a single convolution kernel per layer, dynamic convo

lution aggregates multiple parallel convolution kernels dynamically based upon t heir attentions, which are input dependent. Assembling multiple kernels is not o nly computationally efficient due to the small kernel size, but also has more re presentation power since these kernels are aggregated in a non-linear way via at tention. By simply using dynamic convolution for the state-of-the-art architectu re MobileNetV3-Small, the top-1 accuracy of ImageNet classification is boosted b y 2.9% with only 4% additional FLOPs and 2.9 AP gain is achieved on COCO keypoin t detection.
****

Transform and Tell: Entity-Aware News Image Captioning
Alasdair Tran, Alexander Mathews, Lexing Xie; Proceedings of the IEEE/CVF Conf erence on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13035-13045
We propose an end-to-end model which generates captions for images embedded in n ews articles. News images present two key challenges: they rely on real-world kn owledge, especially about named entities; and they typically have linguistically rich captions that include uncommon words. We address the first challenge by as sociating words in the caption with faces and objects in the image, via a multi-modal, multi-head attention mechanism. We tackle the second challenge with a sta te-of-the-art transformer language model that uses byte-pair-encoding to generat e captions as a sequence of word parts. On the GoodNews dataset, our model outpe rforms the previous state of the art by a factor of four in CIDEr score (13 to 5 4). This performance gain comes from a unique combination of language models, wo rd representation, image embeddings, face embeddings, object embeddings, and imp rovements in neural network design. We also introduce the NYTimes800k dataset wh ich is 70% larger than GoodNews, has higher article quality, and includes the lo cations of images within articles as an additional contextual cue.
****

MTL-NAS: Task-Agnostic Neural Architecture Search Towards General-Purpose Multi-Task Learning
Yuan Gao, Haoping Bai, Zequn Jie, Jiayi Ma, Kui Jia, Wei Liu; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202 0, pp. 11543-11552
We propose to incorporate neural architecture search (NAS) into general-purpose multi-task learning (GP-MTL). Existing NAS methods typically define different se arch spaces according to different tasks. In order to adapt to different task co mbinations (i.e., task sets), we disentangle the GP-MTL networks into single-tas k backbones (optionally encode the task priors), and a hierarchical and layerwis e features sharing/fusing scheme across them. This enables us to design a novel and general task-agnostic search space, which inserts cross-task edges (i.e., fe ature fusion connections) into fixed single-task network backbones. Moreover, we also propose a novel single-shot gradient-based search algorithm that closes th e performance gap between the searched architectures and the final evaluation ar chitecture. This is realized with a minimum entropy regularization on the archit ecture weights during the search phase, which makes the architecture weights con verge to near-discrete values and therefore achieves a single model. As a result , our searched model can be directly used for evaluation without (re-)training f rom scratch. We perform extensive experiments using different single-task backbo nes on various task sets, demonstrating the promising performance obtained by ex ploiting the hierarchical and layerwise features, as well as the desirable gener alizability to different i) task sets and ii) single-task backbones. The code of our paper is available at https://github.com/bhpfelix/MTLNAS.
****

12-in-1: Multi-Task Vision and Language Representation Learning
Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, Stefan Lee; Procee dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP R), 2020, pp. 10437-10446
Much of vision-and-language research focuses on a small but diverse set of indep endent tasks and supporting datasets often studied in isolation; however, the vi sually-grounded language understanding skills required for success at these task s overlap significantly. In this work, we investigate these relationships betwee

n vision-and-language tasks by developing a large-scale, multi-task model. Our a
pproach culminates in a single model on 12 datasets from four broad categories o
f task including visual question answering, caption-based image retrieval, groun
ding referring expressions, and multimodal verification. Compared to independent
ly trained single-task models, this represents a reduction from approximately 3
billion parameters to 270 million while simultaneously improving performance by
2.05 points on average across tasks. We use our multi-task framework to perform
in-depth analysis of the effect of joint training diverse tasks. Further, we sho
w that finetuning task-specific models from our single multi-task model can lead
 to further improvements, achieving performance at or above the state-of-the-art
.
****

Disentangling Physical Dynamics From Unknown Factors for Unsupervised Video Pred
iction

Vincent Le Guen, Nicolas Thome; Proceedings of the IEEE/CVF Conference on Compu
ter Vision and Pattern Recognition (CVPR), 2020, pp. 11474-11484

Leveraging physical knowledge described by partial differential equations (PDEs)
 is an appealing way to improve unsupervised video forecasting models. Since phy
sics is too restrictive for describing the full visual content of generic video
sequences, we introduce PhyDNet, a two-branch deep architecture, which explicitl
y disentangles PDE dynamics from unknown complementary information. A second con
tribution is to propose a new recurrent physical cell (PhyCell), inspired from d
ata assimilation techniques, for performing PDE-constrained prediction in latent
 space. Extensive experiments conducted on four various datasets show the abilit
y of PhyDNet to outperform state-of-the-art methods. Ablation studies also highl
ight the important gain brought out by both disentanglement and PDE-constrained
prediction. Finally, we show that PhyDNet presents interesting features for deal
ing with missing data and long-term forecasting.
****

Gold Seeker: Information Gain From Policy Distributions for Goal-Oriented Vision
-and-Langauge Reasoning

Ehsan Abbasnejad, Iman Abbasnejad, Qi Wu, Javen Shi, Anton van den Hengel; P
roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
 (CVPR), 2020, pp. 13450-13459

As Computer Vision moves from passive analysis of pixels to active analysis of s
emantics, the breadth of information algorithms need to reason over has expanded
 significantly. One of the key challenges in this vein is the ability to identif
y the information required to make a decision, and select an action that will re
cover it. We propose a reinforcement-learning approach that maintains a distribu
tion over its internal information, thus explicitly representing the ambiguity i
n what it knows, and needs to know, towards achieving its goal. Potential action
s are then generated according to this distribution. For each potential action a
 distribution of the expected outcomes is calculated, and the value of the poten
tial information gain assessed. The action taken is that which maximizes the pot
ential information gain. We demonstrate this approach applied to two vision-and-
language problems that have attracted significant recent interest, visual dialog
 and visual query generation. In both cases the method actively selects actions
that will best reduce its internal uncertainty, and outperforms its competitors
in achieving the goal of the challenge.
****

Beyond Short-Term Snippet: Video Relation Detection With Spatio-Temporal Global
Context

Chenchen Liu, Yang Jin, Kehan Xu, Guoqiang Gong, Yadong Mu; Proceedings of t
he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020,
pp. 10840-10849

Video visual relation detection (VidVRD) aims to describe all interacting object
s in a video. Different from relationships in static images, videos contain an a
ddition temporal channel. A majority of existing works divide a video into short
 segments, predict relationships in each segment, and merge them. Such methods c
annot capture relations involving long motions. Predicting the same relationship

across neighboring video segments is also inefficient. To address these issues, this work proposes a novel sliding-window scheme to simultaneously predict short-term and long-term relationships. We run windows with different kernel sizes on object tracklets to generate sub-tracklet proposals with different duration, while the computational load is similar to that in segment-based methods. To fully utilize spatial and temporal information in videos, we construct one spatial and one temporal graph and employ Graph Convloutional Network to generate contextual embedding for tracklet proposal compatibility evaluation. We only predict relationships on highly-compatible proposal pairs. Our method achieves state-of-the-art performance on both ImageNet-VidVRD and VidOR dataset across multiple tasks. Especially for ImageNet-VidVRD, we obtain an average of 3% (R@50 from 8.07% to 11.21%) improvement under all evaluation metrics.

****

Semi-Supervised Semantic Image Segmentation With Self-Correcting Networks

Mostafa S. Ibrahim, Arash Vahdat, Mani Ranjbar, William G. Macready; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12715-12725

Building a large image dataset with high-quality object masks for semantic segmentation is costly and time-consuming. In this paper, we introduce a principled semi-supervised framework that only use a small set of fully supervised images (having semantic segmentation labels and box labels) and a set of images with only object bounding box labels (we call it the weak-set). Our framework trains the primary segmentation model with the aid of an ancillary model that generates initial segmentation labels for the weak-set and a self-correction module that improves the generated labels during training using the increasingly accurate primary model. We introduce two variants of the self-correction module using either linear or convolutional functions. Experiments on the PASCAL VOC 2012 and Cityscape datasets show that our models trained with a small fully supervised set perform similar to, or better than, models trained with a large fully supervised set while requiring 7x less annotation effort.

****

BBN: Bilateral-Branch Network With Cumulative Learning for Long-Tailed Visual Recognition

Boyan Zhou, Quan Cui, Xiu-Shen Wei, Zhao-Min Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9719-9728

Our work focuses on tackling the challenging but natural visual recognition task of long-tailed data distribution (i.e., a few classes occupy most of the data, while most classes have rarely few samples). In the literature, class re-balancing strategies (e.g., re-weighting and re-sampling) are the prominent and effective methods proposed to alleviate the extreme imbalance for dealing with long-tailed problems. In this paper, we firstly discover that these re-balancing methods achieving satisfactory recognition accuracy owe to that they could significantly promote the classifier learning of deep networks. However, at the same time, they will unexpectedly damage the representative ability of the learned deep features to some extent. Therefore, we propose a unified Bilateral-Branch Network (BBN) to take care of both representation learning and classifier learning simultaneously, where each branch does perform its own duty separately. In particular, our BBN model is further equipped with a novel cumulative learning strategy, which is designed to first learn the universal patterns and then pay attention to the tail data gradually. Extensive experiments on four benchmark datasets, including the large-scale iNaturalist ones, justify that the proposed BBN can significantly outperform state-of-the-art methods. Furthermore, validation experiments can demonstrate both our preliminary discovery and effectiveness of tailored designs in BBN for long-tailed problems. Our method won the first place in the iNaturalist 2019 large scale species classification competition, and our code is open-source and available at https://github.com/Megvii-Nanjing/BBN.

****

Sketch Less for More: On-the-Fly Fine-Grained Sketch-Based Image Retrieval

Ayan Kumar Bhunia, Yongxin Yang, Timothy M. Hospedales, Tao Xiang, Yi-Zhe So

ng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn ition (CVPR), 2020, pp. 9779-9788
Fine-grained sketch-based image retrieval (FG-SBIR) addresses the problem of ret rieving a particular photo instance given a user's query sketch. Its widespread applicability is however hindered by the fact that drawing a sketch takes time, and most people struggle to draw a complete and faithful sketch. In this paper, we reformulate the conventional FG-SBIR framework to tackle these challenges, wi th the ultimate goal of retrieving the target photo with the least number of str okes possible. We further propose an on-the-fly design that starts retrieving as soon as the user starts drawing. To accomplish this, we devise a reinforcement learning based cross-modal retrieval framework that directly optimizes rank of t he ground-truth photo over a complete sketch drawing episode. Additionally, we i ntroduce a novel reward scheme that circumvents the problems related to irreleva nt sketch strokes, and thus provides us with a more consistent rank list during the retrieval. We achieve superior early-retrieval efficiency over state-of-the- art methods and alternative baselines on two publicly available fine-grained ske tch retrieval datasets.
****

STINet: Spatio-Temporal-Interactive Network for Pedestrian Detection and Traject ory Prediction
Zhishuai Zhang, Jiyang Gao, Junhua Mao, Yukai Liu, Dragomir Anguelov, Congc ong Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Re cognition (CVPR), 2020, pp. 11346-11355
Detecting pedestrians and predicting future trajectories for them are critical t asks for numerous applications, such as autonomous driving. Previous methods eit her treat the detection and prediction as separate tasks or simply add a traject ory regression head on top of a detector. In this work, we present a novel end-t o-end two-stage network: Spatio-Temporal-Interactive Network (STINet). In additi on to 3D geometry modeling of pedestrians, we model the temporal information for each of the pedestrians. To do so, our method predicts both current and past lo cations in the first stage, so that each pedestrian can be linked across frames and the comprehensive spatio-temporal information can be captured in the second stage. Also, we model the interaction among objects with an interaction graph, t o gather the information among the neighboring objects. Comprehensive experiment s on the Lyft Dataset and the recently released large-scale Waymo Open Dataset f or both object detection and future trajectory prediction validate the effective ness of the proposed method. For the Waymo Open Dataset, we achieve a bird-eyes- view (BEV) detection AP of 80.73 and trajectory prediction average displacement error (ADE) of 33.67cm for pedestrians, which establish the state-of-the-art for both tasks.
****

Intelligent Home 3D: Automatic 3D-House Design From Linguistic Descriptions Only
Qi Chen, Qi Wu, Rui Tang, Yuhan Wang, Shuai Wang, Mingkui Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20 20, pp. 12625-12634
Home design is a complex task that normally requires architects to finish with t heir professional skills and tools. It will be fascinating that if one can produ ce a house plan intuitively without knowing much knowledge about home design and experience of using complex designing tools, for example, via natural language. In this paper, we formulate it as a language conditioned visual content generat ion problem that is further divided into a floor plan generation and an interior texture (such as floor and wall) synthesis task. The only control signal of the generation process is the linguistic expression given by users that describe th e house details. To this end, we propose a House Plan Generative Model (HPGM) th at first translates the language input to a structural graph representation and then predicts the layout of rooms with a Graph Conditioned Layout Prediction Net work (GC-LPN) and generates the interior texture with a Language Conditioned Tex ture GAN (LCT-GAN). With some post-processing, the final product of this task is a 3D house model. To train and evaluate our model, we build the first Text--to- -3D House Model dataset, which will be released at: https:// hidden-link-for-sub

mission.
****
Mask Encoding for Single Shot Instance Segmentation
Rufeng Zhang, Zhi Tian, Chunhua Shen, Mingyu You, Youliang Yan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10226-10235
To date, instance segmentation is dominated by two-stage methods, as pioneered by Mask R-CNN. In contrast, one-stage alternatives cannot compete with Mask R-CNN in mask AP, mainly due to the difficulty of compactly representing masks, making the design of one-stage methods very challenging. In this work, we propose a simple single-shot instance segmentation framework, termed mask encoding based instance segmentation (MEInst). Instead of predicting the two-dimensional mask directly, MEInst distills it into a compact and fixed-dimensional representation vector, which allows the instance segmentation task to be incorporated into one-stage bounding-box detectors and results in a simple yet efficient instance segmentation framework. The proposed one-stage MEInst achieves 36.4% in mask AP with single-model (ResNeXt-101-FPN backbone) and single-scale testing on the MS-COCO benchmark. We show that the much simpler and flexible one-stage instance segmentation method, can also achieve competitive performance. This framework can be easily adapted for other instance-level recognition tasks. Code is available at: git.io/AdelaiDet
****
CentripetalNet: Pursuing High-Quality Keypoint Pairs for Object Detection
Zhiwei Dong, Guoxuan Li, Yue Liao, Fei Wang, Pengju Ren, Chen Qian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10519-10528
Keypoint-based detectors have achieved pretty-well performance. However, incorrect keypoint matching is still widespread and greatly affects the performance of the detector. In this paper, we propose CentripetalNet which uses centripetal shift to pair corner keypoints from the same instance. CentripetalNet predicts the position and the centripetal shift of the corner points and matches corners whose shifted results are aligned. Combining position information, our approach matches corner points more accurately than the conventional embedding approaches do. Corner pooling extracts information inside the bounding boxes onto the border. To make this information more aware at the corners, we design a cross-star deformable convolution network to conduct feature adaption. Furthermore, we explore instance segmentation on anchor-free detectors by equipping our CentripetalNet with a mask prediction module. On COCO test-dev, our CentripetalNet not only outperforms all existing anchor-free detectors with an AP of 48.0% but also achieves comparable performance to the state-of-the-art instance segmentation approaches with a 40.2% Mask AP. Code is available at https: //github.com/KiveeDong/CentripetalNet.
****
Hierarchical Feature Embedding for Attribute Recognition
Jie Yang, Jiarou Fan, Yiru Wang, Yige Wang, Weihao Gan, Lin Liu, Wei Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13055-13064
Attribute recognition is a crucial but challenging task due to viewpoint changes, illumination variations and appearance diversities, etc. Most of previous work only consider the attribute-level feature embedding, which might perform poorly in complicated heterogeneous conditions. To address this problem, we propose a hierarchical feature embedding (HFE) framework, which learns a fine-grained feature embedding by combining attribute and ID information. In HFE, we maintain the inter-class and intra-class feature embedding simultaneously. Not only samples with the same attribute but also samples with the same ID are gathered more closely, which could restrict the feature embedding of visually hard samples with regard to attributes and improve the robustness to variant conditions. We establish this hierarchical structure by utilizing HFE loss consisted of attribute-level and ID-level constraints. We also introduce an absolute boundary regularization and a dynamic loss weight as supplementary components to help build up the feat

ure embedding. Experiments show that our method achieves the state-of-the-art results on two pedestrian attribute datasets and a facial attribute dataset.
****

Mixture Dense Regression for Object Detection and Human Pose Estimation
Ali Varamesh, Tinne Tuytelaars; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13086-13095
Mixture models are well-established learning approaches that, in computer vision, have mostly been applied to inverse or ill-defined problems. However, they are general-purpose divide-and-conquer techniques, splitting the input space into relatively homogeneous subsets in a data-driven manner. Not only ill-defined but also well-defined complex problems should benefit from them. To this end, we devise a framework for spatial regression using mixture density networks. We realize the framework for object detection and human pose estimation. For both tasks, a mixture model yields higher accuracy and divides the input space into interpretable modes. For object detection, mixture components focus on object scale, with the distribution of components closely following that of ground truth the object scale. This practically alleviates the need for multi-scale testing, providing a superior speed-accuracy trade-off. For human pose estimation, a mixture model divides the data based on viewpoint and uncertainty -- namely, front and back views, with back view imposing higher uncertainty. We conduct experiments on the MS COCO dataset and do not face any mode collapse.
****

Don't Even Look Once: Synthesizing Features for Zero-Shot Detection
Pengkai Zhu, Hanxiao Wang, Venkatesh Saligrama; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11693-11702
Zero-shot detection, namely, localizing both seen and unseen objects, increasingly gains importance for large-scale applications, with large number of object classes, since, collecting sufficient annotated data with ground truth bounding boxes is simply not scalable. While vanilla deep neural networks deliver high performance for objects available during training, unseen object detection degrades significantly. At a fundamental level, while vanilla detectors are capable of proposing bounding boxes, which include unseen objects, they are often incapable of assigning high-confidence to unseen objects, due to the inherent precision/recall tradeoffs that requires rejecting background objects. We propose a novel detection algorithm "Don't Even Look Once (DELO)," that synthesizes visual features for unseen objects and augments existing training algorithms to incorporate unseen object detection. Our proposed scheme is evaluated on Pascal VOC and MSCOCO, and we demonstrate significant improvements in test accuracy over vanilla and other state-of-art zero-shot detectors
****

Detection in Crowded Scenes: One Proposal, Multiple Predictions
Xuangeng Chu, Anlin Zheng, Xiangyu Zhang, Jian Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12214-12223
We propose a simple yet effective proposal-based object detector, aiming at detecting highly-overlapped instances in crowded scenes. The key of our approach is to let each proposal predict a set of correlated instances rather than a single one in previous proposal-based frameworks. Equipped with new techniques such as EMD Loss and Set NMS, our detector can effectively handle the difficulty of detecting highly overlapped objects. On a FPN-Res50 baseline, our detector can obtain 4.9% AP gains on challenging CrowdHuman dataset and 1.0% \text MR ^ -2 improvements on CityPersons dataset, without bells and whistles. Moreover, on less crowed datasets like COCO, our approach can still achieve moderate improvement, suggesting the proposed method is robust to crowdedness.
****

Background Data Resampling for Outlier-Aware Classification
Yi Li, Nuno Vasconcelos; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13218-13227
The problem of learning an image classifier that allows detection of out-of-dist

ribution (OOD) examples, with the help of auxiliary background datasets, is studied. While training with background has been shown to improve OOD detection performance, the optimal choice of such dataset remains an open question, and challenges of data imbalance and computational complexity make it a potentially inefficient or even impractical solution. Targeted at balancing between efficiency and detection quality, a dataset resampling approach is proposed for obtaining a compact yet representative set of background data points. The resampling algorithm takes inspiration from prior work on hard negative mining, performing an iterative adversarial weighting on the background examples and using the learned weights to obtain the subset of desired size. Experiments on different datasets, model architectures and training strategies validate the universal effectiveness and efficiency of adversarially resampled background data. Code is available at https://github.com/JerryYLi/ bg-resample-ood.
****

Prime Sample Attention in Object Detection
Yuhang Cao, Kai Chen, Chen Change Loy, Dahua Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11583-11591
It is a common paradigm in object detection frameworks to treat all samples equally and target at maximizing the performance on average. In this work, we revisit this paradigm through a careful study on how different samples contribute to the overall performance measured in terms of mAP. Our study suggests that the samples in each mini-batch are neither independent nor equally important, and therefore a better classifier on average does not necessarily result in higher mAP. Motivated by this study, we propose the notion of Prime Samples, those that play a key role in driving the detection performance. We further develop a simple yet effective sampling and learning strategy called PrIme Sample Attention (PISA) that directs the focus of the training process towards such samples. Our experiments demonstrate that it is often more effective to focus on prime samples than hard samples when training a detector. Particularly, on the MSCOCO dataset, PISA outperforms the random sampling baseline and hard mining schemes, e.g. OHEM and Focal Loss, consistently by around 2% on both single-stage and two-stage detectors, even with a strong backbone ResNeXt-101. Code is available at: https://github.com/open-mmlab/mmdetection.
****

Learning Temporal Co-Attention Models for Unsupervised Video Action Localization
Guoqiang Gong, Xinghan Wang, Yadong Mu, Qi Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9819-9828
Temporal action localization (TAL) in untrimmed videos recently receives tremendous research enthusiasm. To our best knowledge, this is the first attempt in the literature to explore this task under an unsupervised setting, hereafter referred to as action co-localization (ACL), where only the total count of unique actions that appear in the video set is known. To solve ACL, we propose a two-step "clustering + localization" iterative procedure. The clustering step provides noisy pseudo-labels for the localization step, and the localization step provides temporal co-attention models that in turn improve the clustering performance. Using such two-step procedure, weakly-supervised TAL can be regarded as a direct extension of our ACL model. Technically, our contributions are two-folds: 1) temporal co-attention models, either class-specific or class-agnostic, learned from video-level labels or pseudo-labels in an iterative reinforced fashion; 2) new losses specially designed for ACL, including action-background separation loss and cluster-based triplet loss. Comprehensive evaluations are conducted on 20-action THUMOS14 and 100-action ActivityNet-1.2. On both benchmarks, the proposed model for ACL exhibits strong performances, even surprisingly comparable with state-of-the-art weakly-supervised methods. For example, previous best weakly-supervised model achieves 26.8% under mAP@0.5 on THUMOS14, our new records are 30.1% (weakly-supervised) and 25.0% (unsupervised).
****

NAS-FCOS: Fast Neural Architecture Search for Object Detection

Ning Wang, Yang Gao, Hao Chen, Peng Wang, Zhi Tian, Chunhua Shen, Yanning Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11943-11951

The success of deep neural networks relies on significant architecture engineering. Recently neural architecture search (NAS) has emerged as a promise to greatly reduce manual effort in network design by automatically searching for optimal architectures, although typically such algorithms need an excessive amount of computational resources, e.g., a few thousand GPU-days. To date, on challenging vision tasks such as object detection, NAS, especially fast versions of NAS, is less studied. Here we propose to search for the decoder structure of object detectors with search efficiency being taken into consideration. To be more specific, we aim to efficiently search for the feature pyramid network (FPN) as well as the prediction head of a simple anchor-free object detector, namely FCOS, using a tailored reinforcement learning paradigm. With carefully designed search space, search algorithms and strategies for evaluating network quality, we are able to efficiently search a top-performing detection architecture within 4 days using 8 V100 GPUs. The discovered architecture surpasses state-of-the-art object detection models (such as Faster R-CNN, RetinaNet and FCOS) by 1.5 to 3.5 points in AP on the COCO dataset, with comparable computation complexity and memory footprint, demonstrating the efficacy of the proposed NAS for object detection.

****

Enhancing Generic Segmentation With Learned Region Representations

Or Isaacs, Oran Shayer, Michael Lindenbaum; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12946-12955

Deep learning approaches to generic (non-semantic) segmentation have so far been indirect and relied on edge detection. This is in contrast to semantic segmentation, where DNNs are applied directly. We propose an alternative approach called Deep Generic Segmentation (DGS) and try to follow the path used for semantic segmentation. Our main contribution is a new method for learning a pixel-wise representation that reflects segment relatedness. This representation is combined with a CRF to yield the segmentation algorithm. We show that we are able to learn meaningful representations that improve segmentation quality and that the representations themselves achieve state-of-the-art segment similarity scores. The segmentation results are competitive and promising.

****

What's Hidden in a Randomly Weighted Neural Network?

Vivek Ramanujan, Mitchell Wortsman, Aniruddha Kembhavi, Ali Farhadi, Mohammad Rastegari; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11893-11902

Training a neural network is synonymous with learning the values of the weights. By contrast, we demonstrate that randomly weighted neural networks contain subnetworks which achieve impressive performance without ever training the weight values. Hidden in a randomly weighted Wide ResNet-50 is a subnetwork (with random weights) that is smaller than, but matches the performance of a ResNet-34 trained on ImageNet. Not only do these "untrained subnetworks" exist, but we provide an algorithm to effectively find them. We empirically show that as randomly weighted neural networks with fixed weights grow wider and deeper, an "untrained subnetwork" approaches a network with learned weights in accuracy.

****

Learning Texture Invariant Representation for Domain Adaptation of Semantic Segmentation

Myeongjin Kim, Hyeran Byun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12975-12984

Since annotating pixel-level labels for semantic segmentation is laborious, leveraging synthetic data is an attractive solution. However, due to the domain gap between synthetic domain and real domain, it is challenging for a model trained with synthetic data to generalize to real data. In this paper, considering the fundamental difference between the two domains as the texture, we propose a method to adapt to the target domain's texture. First, we diversity the texture of synthetic images using a style transfer algorithm. The various textures of generat

ed images prevent a segmentation model from overfitting to one specific (synthet
ic) texture. Then, we fine-tune the model with self-training to get direct super
vision of the target texture. Our results achieve state-of-the-art performance a
nd we analyze the properties of the model trained on the stylized dataset with e
xtensive experiments.
****

VQA With No Questions-Answers Training
Ben-Zion Vatashsky, Shimon Ullman; Proceedings of the IEEE/CVF Conference on Co
mputer Vision and Pattern Recognition (CVPR), 2020, pp. 10376-10386
Methods for teaching machines to answer visual questions have made significant p
rogress in recent years, but current methods still lack important human capabili
ties, including integrating new visual classes and concepts in a modular manner,
 providing explanations for the answers and handling new domains without explici
t examples. We propose a novel method that consists of two main parts: generatin
g a question graph representation, and an answering procedure, guided by the abs
tract structure of the question graph to invoke an extendable set of visual esti
mators. Training is performed for the language part and the visual part on their
 own, but unlike existing schemes, the method does not require any training usin
g images with associated questions and answers. This approach is able to handle
novel domains (extended question types and new object classes, properties and re
lations) as long as corresponding visual estimators are available. In addition,
it can provide explanations to its answers and suggest alternatives when questio
ns are not grounded in the image. We demonstrate that this approach achieves bot
h high performance and domain extensibility without any questions-answers traini
ng.
****

MCEN: Bridging Cross-Modal Gap between Cooking Recipes and Dish Images with Late
nt Variable Model
Han Fu, Rui Wu, Chenghao Liu, Jianling Sun; Proceedings of the IEEE/CVF Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14570-14580
Nowadays, driven by the increasing concern on diet and health, food computing ha
s attracted enormous attention from both industry and research community. One of
 the most popular research topics in this domain is Food Retrieval, due to its p
rofound influence on health-oriented applications. In this paper, we focus on th
e task of cross-modal retrieval between food images and cooking recipes. We pres
ent Modality-Consistent Embedding Network (MCEN) that learns modality-invariant
representations by projecting images and texts to the same embedding space. To c
apture the latent alignments between modalities, we incorporate stochastic laten
t variables to explicitly exploit the interactions between textual and visual fe
atures. Importantly, our method learns the cross-modal alignments during trainin
g but computes embeddings of different modalities independently at inference tim
e for the sake of efficiency. Extensive experimental results clearly demonstrate
 that the proposed MCEN outperforms all existing approaches on the benchmark Rec
ipe1M dataset and requires less computational cost.
****

NETNet: Neighbor Erasing and Transferring Network for Better Single Shot Object
Detection
Yazhao Li, Yanwei Pang, Jianbing Shen, Jiale Cao, Ling Shao; Proceedings of
the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020,
 pp. 13349-13358
Due to the advantages of real-time detection and improved performance, single-sh
ot detectors have gained great attention recently. To solve the complex scale va
riations, single-shot detectors make scale-aware predictions based on multiple p
yramid layers. However, the features in the pyramid are not scale-aware enough,
which limits the detection performance. Two common problems in single-shot detec
tors caused by object scale variations can be observed: (1) small objects are ea
sily missed; (2) the salient part of a large object is sometimes detected as an
object. With this observation, we propose a new Neighbor Erasing and Transferrin
g (NET) mechanism to reconfigure the pyramid features and explore scale-aware fe
atures. In NET, a Neighbor Erasing Module (NEM) is designed to erase the salient

features of large objects and emphasize the features of small objects in shallow layers. A Neighbor Transferring Module (NTM) is introduced to transfer the erased features and highlight large objects in deep layers. With this mechanism, a single-shot network called NETNet is constructed for scale-aware object detection. In addition, we propose to aggregate nearest neighboring pyramid features to enhance our NET. NETNet achieves 38.5% AP at a speed of 27 FPS and 32.0% AP at a speed of 55 FPS on MS COCO dataset. As a result, NETNet achieves a better trade-off for real-time and accurate object detection.
****

Detailed 2D-3D Joint Representation for Human-Object Interaction
Yong-Lu Li, Xinpeng Liu, Han Lu, Shiyi Wang, Junqi Liu, Jiefeng Li, Cewu Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10166-10175
Human-Object Interaction (HOI) detection lies at the core of action understanding. Besides 2D information such as human/object appearance and locations, 3D pose is also usually utilized in HOI learning since its view-independence. However, rough 3D body joints just carry sparse body information and are not sufficient to understand complex interactions. Thus, we need detailed 3D body shape to go further. Meanwhile, the interacted object in 3D is also not fully studied in HOI learning. In light of these, we propose a detailed 2D-3D joint representation learning method. First, we utilize the single-view human body capture method to obtain detailed 3D body, face and hand shapes. Next, we estimate the 3D object location and size with reference to the 2D human-object spatial configuration and object category priors. Finally, a joint learning framework and cross-modal consistency tasks are proposed to learn the joint HOI representation. To better evaluate the 2D ambiguity processing capacity of models, we propose a new benchmark named Ambiguous-HOI consisting of hard ambiguous images. Extensive experiments in large-scale HOI benchmark and Ambiguous-HOI show impressive effectiveness of our method. Code and data are available at https://github.com/DirtyHarryLYL/DJ-RN.
****

A Programmatic and Semantic Approach to Explaining and Debugging Neural Network Based Object Detectors
Edward Kim, Divya Gopinath, Corina Pasareanu, Sanjit A. Seshia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11128-11137
Even as deep neural networks have become very effective for tasks in vision and perception, it remains difficult to explain and debug their behavior. In this paper, we present a programmatic and semantic approach to explaining, understanding, and debugging the correct and incorrect behaviors of a neural network based perception system. Our approach is semantic in that it employs a high-level representation of the distribution of environment scenarios that the detector is intended to work on. It is programmatic in that the representation is a program in a domain-specific probabilistic programming language using which synthetic data can be generated to train and test the neural network. We present a framework that assesses the performance of the neural network to identify correct and incorrect detections, extracts rules from those results that semantically characterizes the correct and incorrect scenarios, and then specializes the probabilistic program with those rules in order to more precisely characterize the scenarios in which the neural network operates correctly or not, without human intervention. We demonstrate our results using the Scenic probabilistic programming language and a neural network-based object detector. Our experiments show that it is possible to automatically generate compact rules that significantly increase the correct detection rate (or conversely the incorrect detection rate) of the network and can thus help with debugging and understanding its behavior.
****

ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks
Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, Qinghua Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11534-11542
Recently, channel attention mechanism has demonstrated to offer great potential

in improving the performance of deep convolutional neural networks (CNNs). Howev
er, most existing methods dedicate to developing more sophisticated attention mo
dules for achieving better performance, which inevitably increase model complexi
ty. To overcome the paradox of performance and complexity trade-off, this paper
proposes an Efficient Channel Attention (ECA) module, which only involves a hand
ful of parameters while bringing clear performance gain. By dissecting the chann
el attention module in SENet, we empirically show avoiding dimensionality reduct
ion is important for learning channel attention, and appropriate cross-channel i
nteraction can preserve performance while significantly decreasing model complex
ity. Therefore, we propose a local cross-channel interaction strategy without di
mensionality reduction, which can be efficiently implemented via 1D convolution.
Furthermore, we develop a method to adaptively select kernel size of 1D convolu
tion, determining coverage of local cross-channel interaction. The proposed ECA
module is both efficient and effective, e.g., the parameters and computations of
our modules against backbone of ResNet50 are 80 vs. 24.37M and 4.7e-4 GFlops vs
. 3.86 GFlops, respectively, and the performance boost is more than 2% in terms
of Top-1 accuracy. We extensively evaluate our ECA module on image classificatio
n, object detection and instance segmentation with backbones of ResNets and Mobi
leNetV2. The experimental results show our module is more efficient while perfor
ming favorably against its counterparts.
****

Geometry and Learning Co-Supported Normal Estimation for Unstructured Point Clou
d

Haoran Zhou, Honghua Chen, Yidan Feng, Qiong Wang, Jing Qin, Haoran Xie, F
u Lee Wang, Mingqiang Wei, Jun Wang; Proceedings of the IEEE/CVF Conference on
Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13238-13247

In this paper, we propose a normal estimation method for unstructured point clou
d. We observe that geometric estimators commonly focus more on feature preservat
ion but are hard to tune parameters and sensitive to noise, while learning-based
approaches pursue an overall normal estimation accuracy but cannot well handle
challenging regions such as surface edges. This paper presents a novel normal es
timation method, under the co-support of geometric estimator and deep learning.
To lowering the learning difficulty, we first propose to compute a suboptimal in
itial normal at each point by searching for a best fitting patch. Based on the c
omputed normal field, we design a normal-based height map network (NH-Net) to fi
ne-tune the suboptimal normals. Qualitative and quantitative evaluations demonst
rate the clear improvements of our results over both traditional methods and lea
rning-based methods, in terms of estimation accuracy and feature recovery.
****

DR Loss: Improving Object Detection by Distributional Ranking

Qi Qian, Lei Chen, Hao Li, Rong Jin; Proceedings of the IEEE/CVF Conference o
n Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12164-12172

Most of object detection algorithms can be categorized into two classes: two-sta
ge detectors and one-stage detectors. Recently, many efforts have been devoted t
o one-stage detectors for the simple yet effective architecture. Different from
two-stage detectors, one-stage detectors aim to identify foreground objects from
all candidates in a single stage. This architecture is efficient but can suffer
from the imbalance issue with respect to two aspects: the inter-class imbalance
between the number of candidates from foreground and background classes and the
intra-class imbalance in the hardness of background candidates, where only a fe
w candidates are hard to be identified. In this work, we propose a novel distrib
utional ranking (DR) loss to handle the challenge. For each image, we convert th
e classification problem to a ranking problem, which considers pairs of candidat
es within the image, to address the inter-class imbalance problem. Then, we push
the distributions of confidence scores for foreground and background towards th
e decision boundary. After that, we optimize the rank of the expectations of der
ived distributions in lieu of original pairs. Our method not only mitigates the
intra-class imbalance issue in background candidates but also improves the effic
iency for the ranking algorithm. By merely replacing the focal loss in RetinaNet
with the developed DR loss and applying ResNet-101 as the backbone, mAP of the

single-scale test on COCO can be improved from 39.1% to 41.7% without bells and whistles, which demonstrates the effectiveness of the proposed loss function.
****

End-to-End Camera Calibration for Broadcast Videos
Long Sha, Jennifer Hobbs, Panna Felsen, Xinyu Wei, Patrick Lucey, Sujoy Ganguly; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13627-13636
The increasing number of vision-based tracking systems deployed in production have necessitated fast, robust camera calibration. In the domain of sport, the majority of current work focuses on sports where lines and intersections are easy to extract, and appearance is relatively consistent across venues. However, for more challenging sports like basketball, those techniques are not sufficient. In this paper, we propose an end-to-end approach for single moving camera calibration across challenging scenarios in sports. Our method contains three key modules: 1) area-based court segmentation, 2) camera pose estimation with embedded templates, 3) homography prediction via a spatial transform network (STN). All three modules are connected, enabling end-to-end training. We evaluate our method on a new college basketball dataset and demonstrate state of the art performance in variable and dynamic environments. We also validate our method on the World Cup 2014 dataset to show its competitive performance against the state-of-the-art methods. Lastly, we show that our method is two orders of magnitude faster than the previous state of the art on both datasets.
****

Selective Transfer With Reinforced Transfer Network for Partial Domain Adaptation
Zhihong Chen, Chao Chen, Zhaowei Cheng, Boyuan Jiang, Ke Fang, Xinyu Jin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12706-12714
One crucial aspect of partial domain adaptation (PDA) is how to select the relevant source samples in the shared classes for knowledge transfer. Previous PDA methods tackle this problem by re-weighting the source samples based on their high-level information (deep features). However, since the domain shift between source and target domains, only using the deep features for sample selection is defective. We argue that it is more reasonable to additionally exploit the pixel-level information for PDA problem, as the appearance difference between outlier source classes and target classes is significantly large. In this paper, we propose a reinforced transfer network (RTNet), which utilizes both high-level and pixel-level information for PDA problem. Our RTNet is composed of a reinforced data selector (RDS) based on reinforcement learning (RL), which filters out the outlier source samples, and a domain adaptation model which minimizes the domain discrepancy in the shared label space. Specifically, in the RDS, we design a novel reward based on the reconstruct errors of selected source samples on the target generator, which introduces the pixel-level information to guide the learning of RDS. Besides, we develope a state containing high-level information, which used by the RDS for sample selection. The proposed RDS is a general module, which can be easily integrated into existing DA models to make them fit the PDA situation. Extensive experiments indicate that RTNet can achieve state-of-the-art performance for PDA tasks on several benchmark datasets.
****

Neural Head Reenactment with Latent Pose Descriptors
Egor Burkov, Igor Pasechnik, Artur Grigorev, Victor Lempitsky; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13786-13795
We propose a neural head reenactment system, which is driven by a latent pose representation and is capable of predicting the foreground segmentation alongside the RGB image. The latent pose representation is learned as a part of the entire reenactment system, and the learning process is based solely on image reconstruction losses. We show that despite its simplicity, with a large and diverse enough training dataset, such learning successfully decomposes pose from identity. The resulting system can then reproduce mimics of the driving person and, further

more, can perform cross-person reenactment. Additionally, we show that the learned descriptors are useful for other pose-related tasks, such as keypoint prediction and pose-based retrieval.
****

SaccadeNet: A Fast and Accurate Object Detector
Shiyi Lan, Zhou Ren, Yi Wu, Larry S. Davis, Gang Hua; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10397-10406
Object detection is an essential step towards holistic scene understanding. Most existing object detection algorithms attend to certain object areas once and then predict the object locations. However, scientists have revealed that human do not look at the scene in fixed steadiness. Instead, human eyes move around, locating informative parts to understand the object location. This active perceiving movement process is called saccade. In this paper, inspired by such mechanism, we propose a fast and accurate object detector called SaccadeNet. It contains four main modules, the Center Attentive Module, the Corner Attentive Module, the Attention Transitive Module, and the Aggregation Attentive Module, which allows it to attend to different informative object keypoints actively, and predict object locations from coarse to fine. The Corner Attentive Module is used only during training to extract more informative corner features which brings free-lunch performance boost. On the MS COCO dataset, we achieve the performance of 40.4% mAP at 28 FPS and 30.5% mAP at 118 FPS. Among all the real-time object detectors, our SaccadeNet achieves the best detection performance, which demonstrates the effectiveness of the proposed detection mechanism.
****

Learning Augmentation Network via Influence Functions
Donghoon Lee, Hyunsin Park, Trung Pham, Chang D. Yoo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10961-10970
Data augmentation can impact the generalization performance of an image classification model in a significant way. However, it is currently conducted on the basis of trial and error, and its impact on the generalization performance cannot be predicted during training. This paper considers an influence function that predicts how generalization performance, in terms of validation loss, is affected by a particular augmented training sample. The influence function provides an approximation of the change in validation loss without actually comparing the performances that include and exclude the sample in the training process. Based on this function, a differentiable augmentation network is learned to augment an input training sample to reduce validation loss. The augmented sample is fed into the classification network, and its influence is approximated as a function of the parameters of the last fully-connected layer of the classification network. By backpropagating the influence to the augmentation network, the augmentation network parameters are learned. Experimental results on CIFAR-10, CIFAR-100, and ImageNet show that the proposed method provides better generalization performance than conventional data augmentation methods do.
****

Self-Robust 3D Point Recognition via Gather-Vector Guidance
Xiaoyi Dong, Dongdong Chen, Hang Zhou, Gang Hua, Weiming Zhang, Nenghai Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11516-11524
In this paper, we look into the problem of 3D adversary attack, and propose to leverage the internal properties of the point clouds and the adversarial examples to design a new self-robust deep neural network (DNN) based 3D recognition systems. As a matter of fact, on one hand, point clouds are highly structured. Hence for each local part of clean point clouds, it is possible to learn what is it ("part of a bottle") and its relative position ("upper part of a bottle") to the global object center. On the other hand, with the visual quality constraint, 3D adversarial samples often only produce small local perturbations, thus they will roughly keep the original global center but may cause incorrect local relative position estimation. Motivated by these two properties, we use relative position

(dubbed as "gather-vector") as the adversarial indicator and propose a new robust gather module. Equipped with this module, we further propose a new self-robust 3D point recognition network. Through extensive experiments, we demonstrate that the proposed method can improve the robustness of the target attack under the white-box setting significantly. For I-FGSM based attack, our method reduces the attack success rate from 94.37 % to 75.69 %. For C&W based attack, our method reduces the attack success rate more than 40.00 %. Moreover, our method is complementary to other types of defense methods to achieve better defense results.
****

RiFeGAN: Rich Feature Generation for Text-to-Image Synthesis From Prior Knowledge

Jun Cheng, Fuxiang Wu, Yanling Tian, Lei Wang, Dapeng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10911-10920

Text-to-image synthesis is a challenging task that generates realistic images from a textual sequence, which usually contains limited information compared with the corresponding image and so is ambiguous and abstractive. The limited textual information only describes a scene partly, which will complicate the generation with complementing the other details implicitly and lead to low-quality images. To address this problem, we propose a novel rich feature generating text-to-image synthesis, called RiFeGAN, to enrich the given description. In order to provide additional visual details and avoid conflicting, RiFeGAN exploits an attention-based caption matching model to select and refine the compatible candidate captions from prior knowledge. Given enriched captions, RiFeGAN uses self-attentional embedding mixtures to extract features across them effectually and handle the diverging features further. Then it exploits multi-captions attentional generative adversarial networks to synthesize images from those features. The experiments conducted on widely-used datasets show that the models can generate images from enriched captions effectually and improve the results significantly.
****

Unsupervised Model Personalization While Preserving Privacy and Scalability: An Open Problem

Matthias De Lange, Xu Jia, Sarah Parisot, Ales Leonardis, Gregory Slabaugh, Tinne Tuytelaars; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14463-14472

This work investigates the task of unsupervised model personalization, adapted to continually evolving, unlabeled local user images. We consider the practical scenario where a high capacity server interacts with a myriad of resource-limited edge devices, imposing strong requirements on scalability and local data privacy. We aim to address this challenge within the continual learning paradigm and provide a novel Dual User-Adaptation framework (DUA) to explore the problem. This framework flexibly disentangles user-adaptation into model personalization on the server and local data regularization on the user device, with desirable properties regarding scalability and privacy constraints. First, on the server, we introduce incremental learning of task-specific expert models, subsequently aggregated using a concealed unsupervised user prior. Aggregation avoids retraining, whereas the user prior conceals sensitive raw user data, and grants unsupervised adaptation. Second, local user-adaptation incorporates a domain adaptation point of view, adapting regularizing batch normalization parameters to the user data. We explore various empirical user configurations with different priors in categories and a tenfold of transforms for MIT Indoor Scene recognition, and classify numbers in a combined MNIST and SVHN setup. Extensive experiments yield promising results for data-driven local adaptation and elicit user priors for server adaptation to depend on the model rather than user data. Hence, although user-adaptation remains a challenging open problem, the DUA framework formalizes a principled foundation for personalizing both on server and user device, while maintaining privacy and scalability.
****

Learning From Noisy Anchors for One-Stage Object Detection

Hengduo Li, Zuxuan Wu, Chen Zhu, Caiming Xiong, Richard Socher, Larry S. Da

vis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog nition (CVPR), 2020, pp. 10588-10597
State-of-the-art object detectors rely on regressing and classifying an extensiv e list of possible anchors, which are divided into positive and negative samples based on their intersection-over-union (IoU) with corresponding ground-truth ob jects. Such a harsh split conditioned on IoU results in binary labels that are p otentially noisy and challenging for training. In this paper, we propose to miti gate noise incurred by imperfect label assignment such that the contributions of anchors are dynamically determined by a carefully constructed cleanliness score associated with each anchor. Exploring outputs from both regression and classif ication branches, the cleanliness scores, estimated without incurring any additi onal computational overhead, are used not only as soft labels to supervise the t raining of the classification branch but also sample re-weighting factors for im proved localization and classification accuracy. We conduct extensive experiment s on COCO, and demonstrate, among other things, the proposed approach steadily i mproves RetinaNet by 2% with various backbones.
****

Learning Interactions and Relationships Between Movie Characters
Anna Kukleva, Makarand Tapaswi, Ivan Laptev; Proceedings of the IEEE/CVF Confe rence on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9849-9858
Interactions between people are often governed by their relationships. On the fl ip side, social relationships are built upon several interactions. Two strangers are more likely to greet and introduce themselves while becoming friends over t ime. We are fascinated by this interplay between interactions and relationships, and believe that it is an important aspect of understanding social situations. In this work, we propose neural models to learn and jointly predict interactions , relationships, and the pair of characters that are involved. We note that inte ractions are informed by a mixture of visual and dialog cues, and present a mult imodal architecture to extract meaningful information from them. Localizing the pair of interacting characters in video is a time-consuming process, instead, we train our model to learn from clip-level weak labels. We evaluate our models on the MovieGraphs dataset and show the impact of modalities, use of longer tempor al context for predicting relationships, and achieve encouraging performance usi ng weak labels as compared with ground-truth labels. Code is online.
****

MetaIQA: Deep Meta-Learning for No-Reference Image Quality Assessment
Hancheng Zhu, Leida Li, Jinjian Wu, Weisheng Dong, Guangming Shi; Proceeding s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14143-14152
Recently, increasing interest has been drawn in exploiting deep convolutional ne ural networks (DCNNs) for no-reference image quality assessment (NR-IQA). Despit e of the notable success achieved, there is a broad consensus that training DCNN s heavily relies on massive annotated data. Unfortunately, IQA is a typical smal l sample problem. Therefore, most of the existing DCNN-based IQA metrics operate based on pre-trained networks. However, these pre-trained networks are not desi gned for IQA task, leading to generalization problem when evaluating different t ypes of distortions. With this motivation, this paper presents a no-reference IQ A metric based on deep meta-learning. The underlying idea is to learn the meta-k nowledge shared by human when evaluating the quality of images with various dist ortions, which can then be adapted to unknown distortions easily. Specifically, we first collect a number of NR-IQA tasks for different distortions. Then meta-l earning is adopted to learn the prior knowledge shared by diversified distortion s. Finally, the quality prior model is fine-tuned on a target NR-IQA task for qu ickly obtaining the quality model. Extensive experiments demonstrate that the pr oposed metric outperforms the state-of-the-arts by a large margin. Furthermore, the meta-model learned from synthetic distortions can also be easily generalized to authentic distortions, which is highly desired in real-world applications of IQA metrics.
****

RetinaTrack: Online Single Stage Joint Detection and Tracking

Zhichao Lu, Vivek Rathod, Ronny Votel, Jonathan Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14668-14678
Traditionally multi-object tracking and object detection are performed using separate systems with most prior works focusing exclusively on one of these aspects over the other. Tracking systems clearly benefit from having access to accurate detections, however and there is ample evidence in literature that detectors can benefit from tracking which, for example, can help to smooth predictions over time. In this paper we focus on the tracking-by-detection paradigm for autonomous driving where both tasks are mission critical. We propose a conceptually simple and efficient joint model of detection and tracking, called RetinaTrack, which modifies the popular single stage RetinaNet approach such that it is amenable to instance-level embedding training. We show, via evaluations on the Waymo Open Dataset, that we outperform a recent state of the art tracking algorithm while requiring significantly less computation. We believe that our simple yet effective approach can serve as a strong baseline for future work in this area.
****

End-to-End 3D Point Cloud Instance Segmentation Without Detection
Haiyong Jiang, Feilong Yan, Jianfei Cai, Jianmin Zheng, Jun Xiao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12796-12805
3D instance segmentation plays a predominant role in environment perception of robotics and augmented reality. Many deep learning based methods have been presented recently for this task. These methods rely on either a detection branch to propose objects or a grouping step to assemble same-instance points. However, detection based methods do not ensure a consistent instance label for each point, while the grouping step requires parameter-tuning and is computationally expensive. In this paper, we introduce a novel framework to enable end-to-end instance segmentation without detection and a separate step of grouping. The core idea is to convert instance segmentation to a candidate assignment problem. At first, a set of instance candidates is sampled. Then we propose an assignment module for candidate assignment and a suppression module to eliminate redundant candidates. A mapping between instance labels and instance candidates is further sought to construct an instance grouping loss for the network training. Experimental results demonstrate that our method is more effective and efficient than previous approaches.
****

Noise-Aware Fully Webly Supervised Object Detection
Yunhang Shen, Rongrong Ji, Zhiwei Chen, Xiaopeng Hong, Feng Zheng, Jianzhuang Liu, Mingliang Xu, Qi Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11326-11335
We investigate the emerging task of learning object detectors with sole image-level labels on the web without requiring any other supervision like precise annotations or additional images from well-annotated benchmark datasets. Such a task, termed as fully webly supervised object detection, is extremely challenging, since image-level labels on the web are always noisy, leading to poor performance of the learned detectors. In this work, we propose an end-to-end framework to jointly learn webly supervised detectors and reduce the negative impact of noisy labels. Such noise is heterogeneous, which is further categorized into two types, namely background noise and foreground noise. Regarding the background noise, we propose a residual learning structure incorporated with weakly supervised detection, which decomposes background noise and models clean data. To explicitly learn the residual feature between clean data and noisy labels, we further propose a spatially-sensitive entropy criterion, which exploits the conditional distribution of detection results to estimate the confidence of background categories being noise. Regarding the foreground noise, a bagging-mixup learning is introduced, which suppresses foreground noisy signals from incorrectly labelled images, whilst maintaining the diversity of training data. We evaluate the proposed approach on popular benchmark datasets by training detectors on web images, which are retrieved by the corresponding category tags from photo-sharing sites. Extensi

ve experiments show that our method achieves significant improvements over the s
tate-of-the-art methods.
****

PFRL: Pose-Free Reinforcement Learning for 6D Pose Estimation
Jianzhun Shao, Yuhang Jiang, Gu Wang, Zhigang Li, Xiangyang Ji; Proceedings
of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20
20, pp. 11454-11463
6D pose estimation from a single RGB image is a challenging and vital task in co
mputer vision. The current mainstream deep model methods resort to 2D images ann
otated with real-world ground-truth 6D object poses, whose collection is fairly
cumbersome and expensive, even unavailable in many cases. In this work, to get r
id of the burden of 6D annotations, we formulate the 6D pose refinement as a Mar
kov Decision Process and impose on the reinforcement learning approach with only
2D image annotations as weakly-supervised 6D pose information, via a delicate r
eward definition and a composite reinforced optimization method for efficient an
d effective policy training. Experiments on LINEMOD and T-LESS datasets demonstr
ate that our Pose-Free approach is able to achieve state-of-the-art performance
compared with the methods without using real-world ground-truth 6D pose labels.
****

Robust Learning Through Cross-Task Consistency
Amir R. Zamir, Alexander Sax, Nikhil Cheerla, Rohan Suri, Zhangjie Cao, Jit
endra Malik, Leonidas J. Guibas; Proceedings of the IEEE/CVF Conference on Comp
uter Vision and Pattern Recognition (CVPR), 2020, pp. 11197-11206
Visual perception entails solving a wide set of tasks (e.g., object detection, d
epth estimation, etc). The predictions made for different tasks out of one image
are not independent, and therefore, are expected to be 'consistent'. We propose
a flexible and fully computational framework for learning while enforcing Cross
-Task Consistency (X-TAC). The proposed formulation is based on 'inference path
invariance' over an arbitrary graph of prediction domains. We observe that learn
ing with cross-task consistency leads to more accurate predictions, better gener
alization to out-of-distribution samples, and improved sample efficiency. This f
ramework also leads to a powerful unsupervised quantity, called 'Consistency Ene
rgy, based on measuring the intrinsic consistency of the system. Consistency Ene
rgy well correlates with the supervised error (r=0.67), thus it can be employed
as an unsupervised robustness metric as well as for detection of out-of-distribu
tion inputs (AUC=0.99). The evaluations were performed on multiple datasets, inc
luding Taskonomy, Replica, CocoDoom, and ApolloScape.
****

Exploring Self-Attention for Image Recognition
Hengshuang Zhao, Jiaya Jia, Vladlen Koltun; Proceedings of the IEEE/CVF Confer
ence on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10076-10085
Recent work has shown that self-attention can serve as a basic building block fo
r image recognition models. We explore variations of self-attention and assess t
heir effectiveness for image recognition. We consider two forms of self-attentio
n. One is pairwise self-attention, which generalizes standard dot-product attent
ion and is fundamentally a set operator. The other is patchwise self-attention,
which is strictly more powerful than convolution. Our pairwise self-attention ne
tworks match or outperform their convolutional counterparts, and the patchwise m
odels substantially outperform the convolutional baselines. We also conduct expe
riments that probe the robustness of learned representations and conclude that s
elf-attention networks may have significant benefits in terms of robustness and
generalization.
****

Shape correspondence using anisotropic Chebyshev spectral CNNs
Qinsong Li, Shengjun Liu, Ling Hu, Xinru Liu; Proceedings of the IEEE/CVF Con
ference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14658-14667
Establishing correspondence between shapes is a very important and active resear
ch topic in many domains. Due to the powerful ability of deep learning on geomet
ric data, lots of attractive results have been achieved by convolutional neural
networks (CNNs). In this paper, we propose a novel architecture for shape corres

pondence, termed Anisotropic Chebyshev spectral CNNs (ACSCNNs), based on a new e xtension of the manifold convolution operator. The extended convolution operator s aggregate the local features of signals by a set of oriented kernels around ea ch point, which allows to much more comprehensively capture the intrinsic signal information. Rather than using fixed oriented kernels in the spatial domain in previous CNNs, in our framework, the kernels are learned by spectral filtering, based on the eigen-decompositions of multiple Anisotropic Laplace-Beltrami Opera tors. To reduce the computational complexity, we employ an explicit expansion of the Chebyshev polynomial basis to represent the spectral filters whose expansio n coefficients are trainable. Through the benchmark experiments of shape corresp ondence, our architecture is demonstrated to be efficient and be able to provide better than the state-of-the-art results in several datasets even if using cons tant functions as inputs.
****

Uncertainty-Aware Score Distribution Learning for Action Quality Assessment
Yansong Tang, Zanlin Ni, Jiahuan Zhou, Danyang Zhang, Jiwen Lu, Ying Wu, J ie Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern R ecognition (CVPR), 2020, pp. 9839-9848
Assessing action quality from videos has attracted growing attention in recent y ears. Most existing approaches usually tackle this problem based on regression a lgorithms, which ignore the intrinsic ambiguity in the score labels caused by mu ltiple judges or their subjective appraisals. To address this issue, we propose an uncertainty-aware score distribution learning (USDL) approach for action qual ity assessment (AQA). Specifically, we regard an action as an instance associate d with a score distribution, which describes the probability of different evalua ted scores. Moreover, under the circumstance where finer-grained score labels ar e available (e.g., difficulty degree of an action or multiple scores from differ ent judges), we further devise a multi-path uncertainty-aware score distribution learning (MUSDL) method to explore the disentangled components of a score. In o rder to demonstrate the effectiveness of our proposed methods, We conduct experi ments on two AQA datasets containing various Olympic actions. Our approaches set new state-of-the-arts under the Spearman's Rank Correlation (i.e., 0.8102 on AQ A-7 and 0.9273 on MTL-AQA).
****

Seeing Through Fog Without Seeing Fog: Deep Multimodal Sensor Fusion in Unseen A dverse Weather
Mario Bijelic, Tobias Gruber, Fahim Mannan, Florian Kraus, Werner Ritter, K laus Dietmayer, Felix Heide; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11682-11692
The fusion of multimodal sensor streams, such as camera, lidar, and radar measur ements, plays a critical role in object detection for autonomous vehicles, which base their decision making on these inputs. While existing methods exploit redu ndant information in good environmental conditions, they fail in adverse weather where the sensory streams can be asymmetrically distorted. These rare "edge-cas e" scenarios are not represented in available datasets, and existing fusion arch itectures are not designed to handle them. To address this challenge we present a novel multimodal dataset acquired in over 10,000 km of driving in northern Eur ope. Although this dataset is the first large multimodal dataset in adverse weat her, with 100k labels for lidar, camera, radar, and gated NIR sensors, it does n ot facilitate training as extreme weather is rare. To this end, we present a dee p fusion network for robust fusion without a large corpus of labeled training da ta covering all asymmetric distortions. Departing from proposal-level fusion, we propose a single-shot model that adaptively fuses features, driven by measureme nt entropy. We validate the proposed method, trained on clean data, on our exten sive validation dataset. Code and data are available here https://github.com/pri nceton-computational-imaging/SeeingThroughFog.
****

Regularization on Spatio-Temporally Smoothed Feature for Action Recognition
Jinhyung Kim, Seunghwan Cha, Dongyoon Wee, Soonmin Bae, Junmo Kim; Proceedin gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),

2020, pp. 12103-12112
Deep neural networks for video action recognition frequently require 3D convolut ional filters and often encounter overfitting due to a larger number of paramete rs. In this paper, we propose Random Mean Scaling (RMS), a simple and effective regularization method, to relieve the overfitting problem in 3D residual network s. The key idea of RMS is to randomly vary the magnitude of low-frequency compon ents of the feature to regularize the model. The low-frequency component can be derived by a spatio-temporal mean on the local patch of a feature. We present th at selective regularization on this locally smoothed feature makes a model handl e the low-frequency and high-frequency component distinctively, resulting in per formance improvement. RMS can enhance a model with little additional computation only during training, similar to other regularization methods. RMS also can be incorporated into typical training process without any bells and whistles. Exper imental results show the improvement in generalization performance on a popular action recognition datasets demonstrating the effectiveness of RMS as a regulari zation technique, compared to other state-of-the-art regularization methods.
****

Learning Invariant Representation for Unsupervised Image Restoration
Wenchao Du, Hu Chen, Hongyu Yang; Proceedings of the IEEE/CVF Conference on Co mputer Vision and Pattern Recognition (CVPR), 2020, pp. 14483-14492
Recently, cross domain transfer has been applied for unsupervised image restorat ion tasks. However, directly applying existing frameworks would lead to domain-s hift problems in translated images due to lack of effective supervision. Instead , we propose an unsupervised learning method that explicitly learns invariant pr esentation from noisy data and reconstructs clear observations. To do so, we int roduce discrete disentangling representation and adversarial domain adaption int o general domain transfer framework, aided by extra self-supervised modules incl uding background and semantic consistency constraints, learning robust represent ation under dual domain constraints, such as feature and image domains. Experime nts on synthetic and real noise removal tasks show the proposed method achieves comparable performance with other stateof-the-art supervised and unsupervised me thods, while having faster and stable convergence than other domain adaption met hods.
****

Learning Nanoscale Motion Patterns of Vesicles in Living Cells
Arif Ahmed Sekh, Ida Sundvor Opstad, Asa Birna Birgisdottir, Truls Myrmel, B alpreet Singh Ahluwalia, Krishna Agarwal, Dilip K. Prasad; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14014-14023
Detecting and analyzing nanoscale motion patterns of vesicles, smaller than the microscope resolution (250 nm), inside living biological cells is a challenging problem. State-of-the-art CV approaches based on detection, tracking, optical f low or deep learning perform poorly for this problem. We propose an integrative approach, built upon physics based simulations, nanoscopy algorithms, and shallo w residual attention network to make it possible for the first time to analysis sub-resolution motion patterns in vesicles that may also be of sub-resolution di ameter. Our results show state-of-the-art performance, 89% validation accuracy o n simulated dataset and 82% testing accuracy on an experimental dataset of livin g heart muscle cells imaged under three different pathological conditions. We de monstrate automated analysis of the motion states and changed in them for over 9 000 vesicles. Such analysis will enable large scale biological studies of vesicl e transport and interaction in living cells in the future.
****

Network Adjustment: Channel Search Guided by FLOPs Utilization Ratio
Zhengsu Chen, Jianwei Niu, Lingxi Xie, Xuefeng Liu, Longhui Wei, Qi Tian; P roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10658-10667
Automatic designing computationally efficient neural networks has received much attention in recent years. Existing approaches either utilize network pruning or leverage the network architecture search methods. This paper presents a new fra

mework named network adjustment, which considers network accuracy as a function of FLOPs, so that under each network configuration, one can estimate the FLOPs utilization ratio (FUR) for each layer and use it to determine whether to increase or decrease the number of channels on the layer. Note that FUR, like the gradient of a non-linear function, is accurate only in a small neighborhood of the current network. Hence, we design an iterative mechanism so that the initial network undergoes a number of steps, each of which has a small 'adjusting rate' to control the changes to the network. The computational overhead of the entire search process is reasonable, i.e., comparable to that of re-training the final model from scratch. Experiments on standard image classification datasets and a wide range of base networks demonstrate the effectiveness of our approach, which consistently outperforms the pruning counterpart. The code is available at https://github.com/danczs/NetworkAdjustment.

****

StereoGAN: Bridging Synthetic-to-Real Domain Gap by Joint Optimization of Domain Translation and Stereo Matching

Rui Liu, Chengxi Yang, Wenxiu Sun, Xiaogang Wang, Hongsheng Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12757-12766

Large-scale synthetic datasets are beneficial to stereo matching but usually introduce known domain bias. Although unsupervised image-to-image translation networks represented by CycleGAN show great potential in dealing with domain gap, it is non-trivial to generalize this method to stereo matching due to the problem of pixel distortion and stereo mismatch after translation. In this paper, we propose an end-to-end training framework with domain translation and stereo matching networks to tackle this challenge. First, joint optimization between domain translation and stereo matching networks in our end-to-end framework makes the former facilitate the latter one to the maximum extent. Second, this framework introduces two novel losses, i.e., bidirectional multi-scale feature re-projection loss and correlation consistency loss, to help translate all synthetic stereo images into realistic ones as well as maintain epipolar constraints. The effective combination of above two contributions leads to impressive stereo-consistent translation and disparity estimation accuracy. In addition, a mode seeking regularization term is added to endow the synthetic-to-real translation results with higher fine-grained diversity. Extensive experiments demonstrate the effectiveness of the proposed framework on bridging the synthetic-to-real domain gap on stereo matching.

****

Light-weight Calibrator: A Separable Component for Unsupervised Domain Adaptation

Shaokai Ye, Kailu Wu, Mu Zhou, Yunfei Yang, Sia Huat Tan, Kaidi Xu, Jiebo Song, Chenglong Bao, Kaisheng Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13736-13745

Existing domain adaptation methods aim at learning features that can be generalized among domains. These methods commonly require to update source classifier to adapt to the target domain and do not properly handle the trade-off between the source domain and the target domain. In this work, instead of training a classifier to adapt to the target domain, we use a separable component called data calibrator to help the fixed source classifier recover discrimination power in the target domain, while preserving the source domain's performance. When the difference between two domains is small, the source classifier's representation is sufficient to perform well in the target domain and outperforms GAN-based methods in digits. Otherwise, the proposed method can leverage synthetic images generated by GANs to boost performance and achieve state-of-the-art performance in digits datasets and driving scene semantic segmentation. Our method also empirically suggests the potential connection between domain adaptation and adversarial attacks. Code release is available at https://github.com/yeshaokai/ Calibrator-Domain-Adaptation

****

Learning Canonical Shape Space for Category-Level 6D Object Pose and Size Estima

tion
Dengsheng Chen, Jun Li, Zheng Wang, Kai Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11973-11982
We present a novel approach to category-level 6D object pose and size estimation. To tackle intra-class shape variations, we learn canonical shape space (CASS), a unified representation for a large variety of instances of a certain object category. In particular, CASS is modeled as the latent space of a deep generative model of canonical 3D shapes with normalized pose. We train a variational auto-encoder (VAE) for generating 3D point clouds in the canonical space from an RGBD image. The VAE is trained in a cross-category fashion, exploiting the publicly available large 3D shape repositories. Since the 3D point cloud is generated in normalized pose (with actual size), the encoder of the VAE learns view-factorized RGBD embedding. It maps an RGBD image in arbitrary view into a poseindependent 3D shape representation. Object pose is then estimated via contrasting it with a pose-dependent feature of the input RGBD extracted with a separate deep neural networks. We integrate the learning of CASS and pose and size estimation into an end-to-end trainable network, achieving the state-of-the-art performance.
****

A Spatial RNN Codec for End-to-End Image Compression
Chaoyi Lin, Jiabao Yao, Fangdong Chen, Li Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13269-13277
Recently, deep learning has been explored as a promising direction for image compression. Removing the spatial redundancy of the image is crucial for image compression and most learning based methods focus on removing the redundancy between adjacent pixels. Intuitively, to explore larger pixel range beyond adjacent pixel is beneficial for removing the redundancy. In this paper, we propose a fast yet effective method for end-to-end image compression by incorporating a novel spatial recurrent neural network. Block based LSTM is utilized to remove the redundant information between adjacent pixels and blocks. Besides, the proposed method is a potential efficient system that parallel computation on individual blocks is possible. Experimental results demonstrate that the proposed model outperforms state-of-the-art traditional image compression standards and learning based image compression models in terms of both PSNR and MS-SSIM metrics. It provides a 26.73% bits-reduction than High Efficiency Video Coding (HEVC), which is the current official state-of-the-art video codec.
****

Two Causal Principles for Improving Visual Dialog
Jiaxin Qi, Yulei Niu, Jianqiang Huang, Hanwang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10860-10869
This paper unravels the design tricks adopted by us, the champion team MReaL-BDAI, for Visual Dialog Challenge 2019: two causal principles for improving Visual Dialog (VisDial). By "improving", we mean that they can promote almost every existing VisDial model to the state-of-the-art performance on the leader-board. Such a major improvement is only due to our careful inspection on the causality behind the model and data, finding that the community has overlooked two causalities in VisDial. Intuitively, Principle 1 suggests: we should remove the direct input of the dialog history to the answer model, otherwise a harmful shortcut bias will be introduced; Principle 2 says: there is an unobserved confounder for history, question, and answer, leading to spurious correlations from training data. In particular, to remove the confounder suggested in Principle 2, we propose several causal intervention algorithms, which make the training fundamentally different from the traditional likelihood estimation. Note that the two principles are model-agnostic, so they are applicable in any VisDial model.
****

ILFO: Adversarial Attack on Adaptive Neural Networks
Mirazul Haque, Anki Chauhan, Cong Liu, Wei Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14264-14273

With the increasing number of layers and parameters in neural networks, the energy consumption of neural networks has become a great concern to society, especially to users of handheld or embedded devices. In this paper, we investigate the robustness of neural networks against energy-oriented attacks. Specifically, we propose ILFO (Intermediate Output-Based Loss Function Optimization) attack against a common type of energy-saving neural networks, Adaptive Neural Networks (AdNN). AdNNs save energy consumption by dynamically deactivating part of its model based on the need of the inputs. ILFO leverages intermediate output as a proxy to infer the relation between input and its corresponding energy consumption. ILFO has shown an increase up to 100 % of the FLOPs (floating-point operations per second) reduced by AdNNs with minimum noise added to input images. To our knowledge, this is the first attempt to attack the energy consumption of an AdNN.
****

Learning to Evaluate Perception Models Using Planner-Centric Metrics
Jonah Philion, Amlan Kar, Sanja Fidler; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14055-14064
Variants of accuracy and precision are the gold-standard by which the computer vision community measures progress of perception algorithms. One reason for the ubiquity of these metrics is that they are largely task-agnostic; we in general seek to detect zero false negatives or positives. The downside of these metrics is that, at worst, they penalize all incorrect detections equally without conditioning on the task or scene, and at best, heuristics need to be chosen to ensure that different mistakes count differently. In this paper, we propose a principled metric for 3D object detection specifically for the task of self-driving. The core idea behind our metric is to isolate the task of object detection and measure the impact the produced detections would induce on the downstream task of driving. Without hand-designing it to, we find that our metric penalizes many of the mistakes that other metrics penalize by design. In addition, our metric downweighs detections based on additional factors such as distance from a detection to the ego car and the speed of the detection in intuitive ways that other detection metrics do not. For human evaluation, we generate scenes in which standard metrics and our metric disagree and find that humans side with our metric 79% of the time. Our project page including an evaluation server can be found at https://nv-tlabs.github.io/detection-relevance.
****

Hierarchical Clustering With Hard-Batch Triplet Loss for Person Re-Identification
Kaiwei Zeng, Munan Ning, Yaohua Wang, Yang Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13657-13665
For clustering-guided fully unsupervised person reidentification (re-ID) methods, the quality of pseudo labels generated by clustering directly decides the model performance. In order to improve the quality of pseudo labels in existing methods, we propose the HCT method which combines hierarchical clustering with hard-batch triplet loss. The key idea of HCT is to make full use of the similarity among samples in the target dataset through hierarchical clustering, reduce the influence of hard examples through hard-batch triplet loss, so as to generate high quality pseudo labels and improve model performance. Specifically, (1) we use hierarchical clustering to generate pseudo labels, (2) we use PK sampling in each iteration to generate a new dataset for training, (3) we conduct training with hard-batch triplet loss and evaluate model performance in each iteration. We evaluate our model on Market-1501 and DukeMTMC-reID. Results show that HCT achieves 56.4% mAP on Market-1501 and 50.7% mAP on DukeMTMC-reID which surpasses state-of-the-arts a lot in fully unsupervised re-ID and even better than most unsupervised domain adaptation (UDA) methods which use the labeled source dataset. Code will be released soon on https://github.com/zengkaiwei/HCT
****

Fast Template Matching and Update for Video Object Tracking and Segmentation
Mingjie Sun, Jimin Xiao, Eng Gee Lim, Bingfeng Zhang, Yao Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20

20, pp. 10791-10799
In this paper, the main task we aim to tackle is the multi-instance semi-supervised video object segmentation across a sequence of frames where only the first-frame box-level ground-truth is provided. Detection-based algorithms are widely adopted to handle this task, and the challenges lie in the selection of the matching method to predict the result as well as to decide whether to update the target template using the newly predicted result. The existing methods, however, make these selections in a rough and inflexible way, compromising their performance. To overcome this limitation, we propose a novel approach which utilizes reinforcement learning to make these two decisions at the same time. Specifically, the reinforcement learning agent learns to decide whether to update the target template according to the quality of the predicted result. The choice of the matching method will be determined at the same time, based on the action history of the reinforcement learning agent. Experiments show that our method is almost 10 times faster than the previous state-of-the-art method with even higher accuracy (region similarity of 69.1% on DAVIS 2017 dataset).
****

TCTS: A Task-Consistent Two-Stage Framework for Person Search
Cheng Wang, Bingpeng Ma, Hong Chang, Shiguang Shan, Xilin Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11952-11961
The state of the art person search methods separate person search into detection and re-ID stages, but ignore the consistency between these two stages. The general person detector has no special attention on the query target; The re-ID model is trained on hand-drawn bounding boxes which are not available in person search. To address the consistency problem, we introduce a Task-Consist Two-Stage (TCTS) person search framework, includes an identity-guided query (IDGQ) detector and a Detection Results Adapted (DRA) re-ID model. In the detection stage, the IDGQ detector learns an auxiliary identity branch to compute query similarity scores for proposals. With consideration of the query similarity scores and foreground score, IDGQ produces query-like bounding boxes for the re-ID stage. In the re-ID stage, we predict identity labels of detected bounding boxes, and use these examples to construct a more practical mixed train set for the DRA model. Training on the mixed train set improves the robustness of the re-ID stage to inaccurate detection. We evaluate our method on two benchmark datasets, CUHK-SYSU and PRW. Our framework achieves 93.9% of mAP and 95.1% of rank1 accuracy on CUHK-SYSU, outperforming the previous state of the art methods.
****

Cross-domain Object Detection through Coarse-to-Fine Feature Adaptation
Yangtao Zheng, Di Huang, Songtao Liu, Yunhong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13766-13775
Recent years have witnessed great progress in deep learning based object detection. However, due to the domain shift problem, applying off-the-shelf detectors to an unseen domain leads to significant performance drop. To address such an issue, this paper proposes a novel coarse-to-fine feature adaptation approach to cross-domain object detection. At the coarse-grained stage, different from the rough image-level or instance-level feature alignment used in the literature, foreground regions are extracted by adopting the attention mechanism, and aligned according to their marginal distributions via multi-layer adversarial learning in the common feature space. At the fine-grained stage, we conduct conditional distribution alignment of foregrounds by minimizing the distance of global prototypes with the same category but from different domains. Thanks to this coarse-to-fine feature adaptation, domain knowledge in foreground regions can be effectively transferred. Extensive experiments are carried out in various cross-domain detection scenarios. The results are state-of-the-art, which demonstrate the broad applicability and effectiveness of the proposed approach.
****

Efficient Derivative Computation for Cumulative B-Splines on Lie Groups
Christiane Sommer, Vladyslav Usenko, David Schubert, Nikolaus Demmel, Daniel

Cremers; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11148-11156

Continuous-time trajectory representation has recently gained popularity for tasks where the fusion of high-frame-rate sensors and multiple unsynchronized devices is required. Lie group cumulative B-splines are a popular way of representing continuous trajectories without singularities. They have been used in near real-time SLAM and odometry systems with IMU, LiDAR, regular, RGB-D and event cameras, as well as for offline calibration. These applications require efficient computation of time derivatives (velocity, acceleration), but all prior works rely on a computationally suboptimal formulation. In this work we present an alternative derivation of time derivatives based on recurrence relations that needs O(k) instead of O(k^2) matrix operations (for a spline of order k) and results in simple and elegant expressions. While producing the same result, the proposed approach significantly speeds up the trajectory optimization and allows for computing simple analytic derivatives with respect to spline knots. The results presented in this paper pave the way for incorporating continuous-time trajectory representations into more applications where real-time performance is required.

****

Counterfactual Vision and Language Learning

Ehsan Abbasnejad, Damien Teney, Amin Parvaneh, Javen Shi, Anton van den Hengel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10044-10054

The ongoing success of visual question answering methods has been somwehat surprising given that, at its most general, the problem requires understanding the entire variety of both visual and language stimuli. It is particularly remarkable that this success has been achieved on the basis of comparatively small datasets, given the scale of the problem. One explanation is that this has been accomplished partly by exploiting bias in the datasets rather than developing deeper multi-modal reasoning. This fundamentally limits the generalization of the method, and thus its practical applicability. We propose a method that addresses this problem by introducing counterfactuals in the training. In doing so we leverage structural causal models for counterfactual evaluation to formulate alternatives, for instance, questions that could be asked of the same image set. We show that simulating plausible alternative training data through this process results in better generalization.

****

Unsupervised Reinforcement Learning of Transferable Meta-Skills for Embodied Navigation

Juncheng Li, Xin Wang, Siliang Tang, Haizhou Shi, Fei Wu, Yueting Zhuang, William Yang Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12123-12132

Visual navigation is a task of training an embodied agent by intelligently navigating to a target object (e.g., television) using only visual observations. A key challenge for current deep reinforcement learning models lies in the requirements for a large amount of training data. It is exceedingly expensive to construct sufficient 3D synthetic environments annotated with the target object information. In this paper, we focus on visual navigation in the low-resource setting, where we have only a few training environments annotated with object information. We propose a novel unsupervised reinforcement learning approach to learn transferable meta-skills (e.g., bypass obstacles, go straight) from unannotated environments without any supervisory signals. The agent can then fast adapt to visual navigation through learning a high-level master policy to combine these meta-skills, when the visual-navigation-specified reward is provided. Experimental results show that our method significantly outperforms the baseline by 53.34% relatively on SPL, and further qualitative analysis demonstrates that our method learns transferable motor primitives for visual navigation.

****

M2m: Imbalanced Classification via Major-to-Minor Translation

Jaehyung Kim, Jongheon Jeong, Jinwoo Shin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13896-13905

In most real-world scenarios, labeled training datasets are highly class-imbalanced, where deep neural networks suffer from generalizing to a balanced testing criterion. In this paper, we explore a novel yet simple way to alleviate this issue by augmenting less-frequent classes via translating samples (e.g., images) from more-frequent classes. This simple approach enables a classifier to learn more generalizable features of minority classes, by transferring and leveraging the diversity of the majority information. Our experimental results on a variety of class-imbalanced datasets show that the proposed method improves the generalization on minority classes significantly compared to other existing re-sampling or re-weighting methods. The performance of our method even surpasses those of previous state-of-the-art methods for the imbalanced classification.

****

DSGN: Deep Stereo Geometry Network for 3D Object Detection

Yilun Chen, Shu Liu, Xiaoyong Shen, Jiaya Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12536-12545

Most state-of-the-art 3D object detectors rely heavily on LiDAR sensors and there remains a large gap in terms of performance between image-based and LiDAR-based methods, caused by inappropriate representation for the prediction in 3D scenarios. Our method, called Deep Stereo Geometry Network (DSGN), reduces this gap significantly by detecting 3D objects on a differentiable volumetric representation -- 3D geometric volume, which effectively encodes 3D geometric structure for 3D regular space. With this representation, we learn depth information and semantic cues simultaneously. For the first time, we provide a simple and effective one-stage stereo-based 3D detection pipeline that jointly estimates the depth and detects 3D objects in an end-to-end learning manner. Our approach outperforms previous stereo-based 3D detectors (about 10 higher in terms of AP) and even achieves comparable performance with a few LiDAR-based methods on the KITTI 3D object detection leaderboard. Code will be made publicly available at https://github.com/chenyilun95/DSGN.

****

Predicting Semantic Map Representations From Images Using Pyramid Occupancy Networks

Thomas Roddick, Roberto Cipolla; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11138-11147

Autonomous vehicles commonly rely on highly detailed birds-eye-view maps of their environment, which capture both static elements of the scene such as road layout as well as dynamic elements such as other cars and pedestrians. Generating these map representations on the fly is a complex multi-stage process which incorporates many important vision-based elements, including ground plane estimation, road segmentation and 3D object detection. In this work we present a simple, unified approach for estimating these map representations directly from monocular images using a single end-to-end deep learning architecture. For the maps themselves we adopt a semantic Bayesian occupancy grid framework, allowing us to trivially accumulate information over multiple cameras and timesteps. We demonstrate the effectiveness of our approach by evaluating against several challenging baselines on the NuScenes and Argoverse datasets, and show that we are able to achieve a relative improvement of 9.1% and 22.3% respectively compared to the best-performing existing method.

****

Memory Aggregation Networks for Efficient Interactive Video Object Segmentation

Jiaxu Miao, Yunchao Wei, Yi Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10366-10375

Interactive video object segmentation (iVOS) aims at efficiently harvesting high-quality segmentation masks of the target object in a video with user interactions. Most previous state-of-the-arts tackle the iVOS with two independent networks for conducting user interaction and temporal propagation, respectively, leading to inefficiencies during the inference stage. In this work, we propose a unified framework, named Memory Aggregation Networks (MA-Net), to address the challenging iVOS in a more efficient way. Our MA-Net integrates the interaction and the

propagation operations into a single network, which significantly promotes the efficiency of iVOS in the scheme of multi-round interactions. More importantly, we propose a simple yet effective memory aggregation mechanism to record the informative knowledge from the previous interaction rounds, improving the robustness in discovering challenging objects of interest greatly. We conduct extensive experiments on the validation set of DAVIS Challenge 2018 benchmark. In particular, our MA-Net achieves the J@60 score of 76.1% without any bells and whistles, outperforming the state-of-the-arts with more than 2.7%.

****

SegGCN: Efficient 3D Point Cloud Segmentation With Fuzzy Spherical Kernel

Huan Lei, Naveed Akhtar, Ajmal Mian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11611-11620

Fuzzy clustering is known to perform well in real-world applications. Inspired by this observation, we incorporate a fuzzy mechanism into discrete convolutional kernels for 3D point clouds as our first major contribution. The proposed fuzzy kernel is defined over a spherical volume that uses discrete bins. Discrete volumetric division can normally make a kernel vulnerable to boundary effects during learning as well as point density during inference. However, the proposed kernel remains robust to boundary conditions and point density due to the fuzzy mechanism. Our second major contribution comes as the proposal of an efficient graph convolutional network, SegGCN for segmenting point clouds. The proposed network exploits ResNet like blocks in the encoder and 1 x 1 convolutions in the decoder. SegGCN capitalizes on the separable convolution operation of the proposed fuzzy kernel for efficiency. We establish the effectiveness of the SegGCN with the proposed kernel on the challenging S3DIS and ScanNet real-world datasets. Our experiments demonstrate that the proposed network can segment over one million points per second with highly competitive performance.

****

AutoTrack: Towards High-Performance Visual Tracking for UAV With Automatic Spatio-Temporal Regularization

Yiming Li, Changhong Fu, Fangqiang Ding, Ziyuan Huang, Geng Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11923-11932

Most existing trackers based on discriminative correlation filters (DCF) try to introduce predefined regularization term to improve the learning of target objects, e.g., by suppressing background learning or by restricting change rate of correlation filters. However, predefined parameters introduce much effort in tuning them and they still fail to adapt to new situations that the designer did not think of. In this work, a novel approach is proposed to online automatically and adaptively learn spatio-temporal regularization term. Spatially local response map variation is introduced as spatial regularization to make DCF focus on the learning of trust-worthy parts of the object, and global response map variation determines the updating rate of the filter. Extensive experiments on four UAV benchmarks have proven the superiority of our method compared to the state-of-the-art CPU- and GPU-based trackers, with a speed of 60 frames per second running on a single CPU. Our tracker is additionally proposed to be applied in UAV localization. Considerable tests in the indoor practical scenarios have proven the effectiveness and versatility of our localization method. The code is available at https://github.com/vision4robotics/AutoTrack.

****

Multi-Mutual Consistency Induced Transfer Subspace Learning for Human Motion Segmentation

Tao Zhou, Huazhu Fu, Chen Gong, Jianbing Shen, Ling Shao, Fatih Porikli; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10277-10286

Human motion segmentation based on transfer subspace learning is a rising interest in action-related tasks. Although progress has been made, there are still several issues within the existing methods. First, existing methods transfer knowledge from source data to target tasks by learning domain-invariant features, but they ignore to preserve domain-specific knowledge. Second, the transfer subspace

learning is employed in either low-level or high-level feature spaces, but few methods consider fusing multi-level features for subspace learning. To this end, we propose a novel multi-mutual consistency induced transfer subspace learning framework for human motion segmentation. Specifically, our model factorizes the source and target data into distinct multi-layer feature spaces and reduces the distribution gap between them through a multi-mutual consistency learning strategy. In this way, the domain-specific knowledge and domain-invariant properties can be explored simultaneously. Our model also conducts the transfer subspace learning on different layers to capture multi-level structural information. Further, to preserve the temporal correlations, we project the learned representations into a block-like space. The proposed model is efficiently optimized by using the Augmented Lagrange Multiplier (ALM) algorithm. Experimental results on four human motion datasets demonstrate the effectiveness of our method over other state-of-the-art approaches.

****

Associate-3Ddet: Perceptual-to-Conceptual Association for 3D Point Cloud Object Detection

Liang Du, Xiaoqing Ye, Xiao Tan, Jianfeng Feng, Zhenbo Xu, Errui Ding, Shilei Wen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13329-13338

Object detection from 3D point clouds remains a challenging task, though recent studies pushed the envelope with the deep learning techniques. Owing to the severe spatial occlusion and inherent variance of point density with the distance to sensors, appearance of a same object varies a lot in point cloud data. Designing robust feature representation against such appearance changes is hence the key issue in a 3D object detection method. In this paper, we innovatively propose a domain adaptation like approach to enhance the robustness of the feature representation. More specifically, we bridge the gap between the perceptual domain where the feature comes from a real scene and the conceptual domain where the feature is extracted from an augmented scene consisting of non-occlusion point cloud rich of detailed information. This domain adaptation approach mimics the functionality of the human brain when proceeding object perception. Extensive experiments demonstrate that our simple yet effective approach fundamentally boosts the performance of 3D point cloud object detection and achieves the state-of-the-art results.

****

Training a Steerable CNN for Guidewire Detection

Donghang Li, Adrian Barbu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13955-13963

Guidewires are thin wires used in coronary angioplasty to guide different tools to access and repair the obstructed artery. The whole procedure is monitored using fluoroscopic (real-time X-ray) images. Due to the guidewire being thin in the low quality fluoroscopic images, it is usually poorly visible. The poor quality of the X-ray images makes the guidewire detection a challenging problem in image-guided interventions. Localizing the guidewire could help in enhancing its visibility and for other automatic procedures. Guidewire localization methods usually contain a first step of computing a pixelwise guidewire response map on the entire image. In this paper, we present a steerable Convolutional Neural Network (CNN), which is a Fully Convolutional Neural Network (FCNN) that can detect objects rotated by an arbitrary 2D angle, without being rotation invariant. In fact, the steerable CNN has an angle parameter that can be changed to make it sensitive to objects rotated by that angle. We present an application of this idea to detecting the guidewire pixels, and compare it with an FCNN trained to be invariant to the guidewire orientation. Results reveal that the proposed method is a good choice, outperforming some popular filter-based and learning-based approaches such as Frangi Filter, Spherical Quadrature Filter, FCNN and a state of the art trained classifier based on hand-crafted feature.

****

GIFnets: Differentiable GIF Encoding Framework

Innfarn Yoo, Xiyang Luo, Yilin Wang, Feng Yang, Peyman Milanfar; Proceedings

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14473-14482
Graphics Interchange Format (GIF) is a widely used image file format. Due to the limited number of palette colors, GIF encoding often introduces color banding artifacts. Traditionally, dithering is applied to reduce color banding, but introducing dotted-pattern artifacts. To reduce artifacts and provide a better and more efficient GIF encoding, we introduce a differentiable GIF encoding pipeline, which includes three novel neural networks: PaletteNet, DitherNet, and BandingNet. Each of these three networks provides an important functionality within the GIF encoding pipeline. PaletteNet predicts a near-optimal color palette given an input image. DitherNet manipulates the input image to reduce color banding artifacts and provides an alternative to traditional dithering. Finally, BandingNet is designed to detect color banding, and provides a new perceptual loss specifically for GIF images. As far as we know, this is the first fully differentiable GIF encoding pipeline based on deep neural networks and compatible with existing GIF decoders. User study shows that our algorithm is better than Floyd-Steinberg based GIF encoding.
****

TRPLP - Trifocal Relative Pose From Lines at Points
Ricardo Fabbri, Timothy Duff, Hongyi Fan, Margaret H. Regan, David da Costa de Pinho, Elias Tsigaridas, Charles W. Wampler, Jonathan D. Hauenstein, Peter J. Giblin, Benjamin Kimia, Anton Leykin, Tomas Pajdla; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12073-12083
We present a method for solving two minimal problems for relative camera pose estimation from three views, which are based on three view correspondences of (i) three points and one line and (ii) three points and two lines through two of the points. These problems are too difficult to be efficiently solved by the state of the art Grobner basis methods. Our method is based on a new efficient homotopy continuation (HC) solver, which dramatically speeds up previous HC solving by specializing HC methods to generic cases of our problems. We show in simulated experiments that our solvers are numerically robust and stable under image noise. We show in real experiment that (i) SIFT features provide good enough point-and-line correspondences for three-view reconstruction and (ii) that we can solve difficult cases with too few or too noisy tentative matches where the state of the art structure from motion initialization fails.
****

SP-NAS: Serial-to-Parallel Backbone Search for Object Detection
Chenhan Jiang, Hang Xu, Wei Zhang, Xiaodan Liang, Zhenguo Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11863-11872
Advanced object detectors usually adopt a backbone network designed and pretrained by ImageNet classification. Recently neural architecture search (NAS) has emerged to automatically design a task-specific backbone to bridge the gap between the tasks of classification and detection. In this paper, we propose a two-phase serial-to-parallel architecture search framework named SP-NAS towards a flexible task-oriented detection backbone. Specifically, the serial-searching round aims at finding a sequence of serial blocks with optimal scale and output channels in the feature hierarchy by a Swap-Expand-Reignite search algorithm; the parallel-searching phase then assembles several sub-architectures along with the previous searched backbone into a more powerful parallel-structured backbone. We efficiently search a detection backbone by exploring a network morphism strategy on multiple detection benchmarks. The resulting architectures achieve SOTA results, i.e. top performance (LAMR: 0.055) on the automotive detection leaderboard of EuroCityPersons benchmark, improving 2.3% mAP with less FLOPS than NAS-FPN on COCO, and reaching 84.1% AP50 on VOC better than DetNAS and Auto-FPN in terms of both accuracy and speed.
****

Rethinking Depthwise Separable Convolutions: How Intra-Kernel Correlations Lead to Improved MobileNets

Daniel Haase, Manuel Amthor; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14600-14609

We introduce blueprint separable convolutions (BSConv) as highly efficient building blocks for CNNs. They are motivated by quantitative analyses of kernel properties from trained models, which show the dominance of correlations along the depth axis. Based on our findings, we formulate a theoretical foundation from which we derive efficient implementations using only standard layers. Moreover, our approach provides a thorough theoretical derivation, interpretation, and justification for the application of depthwise separable convolutions (DSCs) in general, which have become the basis of many modern network architectures. Ultimately, we reveal that DSC-based architectures such as MobileNets implicitly rely on cross-kernel correlations, while our BSConv formulation is based on intra-kernel correlations and thus allows for a more efficient separation of regular convolutions. Extensive experiments on large-scale and fine-grained classification datasets show that BSConvs clearly and consistently improve MobileNets and other DSC-based architectures without introducing any further complexity. For fine-grained datasets, we achieve an improvement of up to 13.7 percentage points. In addition, if used as drop-in replacement for standard architectures such as ResNets, BSConv variants also outperform their vanilla counterparts by up to 9.5 percentage points on ImageNet.

****

Vision-Dialog Navigation by Exploring Cross-Modal Memory

Yi Zhu, Fengda Zhu, Zhaohuan Zhan, Bingqian Lin, Jianbin Jiao, Xiaojun Chang, Xiaodan Liang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10730-10739

Vision-dialog navigation posed as a new holy-grail task in vision-language disciplinary targets at learning an agent endowed with the capability of constant conversation for help with natural language and navigating according to human responses. Besides the common challenges faced in visual language navigation, vision-dialog navigation also requires to handle well with the language intentions of a series of questions about the temporal context from dialogue history and co-reasoning both dialogs and visual scenes. In this paper, we propose the Cross-modal Memory Network (CMN) for remembering and understanding the rich information relevant to historical navigation actions. Our CMN consists of two memory modules, the language memory module (L-mem) and the visual memory module (V-mem). Specifically, L-mem learns latent relationships between the current language interaction and a dialog history by employing a multi-head attention mechanism. V-mem learns to associate the current visual views and the cross-modal memory about the previous navigation actions. The cross-modal memory is generated via a vision-to-language attention and a language-to-vision attention. Benefiting from the collaborative learning of the L-mem and the V-mem, our CMN is able to explore the memory about the decision making of historical navigation actions which is for the current step. Experiments on the CVDN dataset show that our CMN outperforms the previous state-of-the-art model by a significant margin on both seen and unseen environments.

****

PointRend: Image Segmentation As Rendering

Alexander Kirillov, Yuxin Wu, Kaiming He, Ross Girshick; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9799-9808

We present a new method for efficient high-quality image segmentation of objects and scenes. By analogizing classical computer graphics methods for efficient rendering with over- and undersampling challenges faced in pixel labeling tasks, we develop a unique perspective of image segmentation as a rendering problem. From this vantage, we present the PointRend (Point-based Rendering) neural network module: a module that performs point-based segmentation predictions at adaptively selected locations based on an iterative subdivision algorithm. PointRend can be flexibly applied to both instance and semantic segmentation tasks by building on top of existing state-of-the-art models. While many concrete implementations of the general idea are possible, we show that a simple design already achieves

excellent results. Qualitatively, PointRend outputs crisp object boundaries in regions that are over-smoothed by previous methods. Quantitatively, PointRend yields significant gains on COCO and Cityscapes, for both instance and semantic segmentation. PointRend's efficiency enables output resolutions that are otherwise impractical in terms of memory or computation compared to existing approaches. Code has been made available at https://github.com/facebookresearch/detectron2/tree/master/projects/PointRend.

****

Differentiable Adaptive Computation Time for Visual Reasoning

Cristobal Eyzaguirre, Alvaro Soto; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12817-12825

This paper presents a novel attention-based algorithm for achieving adaptive computation called DACT, which, unlike existing ones, is end-to-end differentiable. Our method can be used in conjunction with many networks; in particular, we study its application to the widely know MAC architecture, obtaining a significant reduction in the number of recurrent steps needed to achieve similar accuracies, therefore improving its performance to computation ratio. Furthermore, we show that by increasing the maximum number of steps used, we surpass the accuracy of even our best non-adaptive MAC in the CLEVR dataset, demonstrating that our approach is able to control the number of steps without significant loss of performance. Additional advantages provided by our approach include considerably improving interpretability by discarding useless steps and providing more insights into the underlying reasoning process. Finally, we present adaptive computation as an equivalent to an ensemble of models, similar to a mixture of expert formulation. Both the code and the configuration files for our experiments are made available to support further research in this area.

****

Exploring Data Aggregation in Policy Learning for Vision-Based Urban Autonomous Driving

Aditya Prakash, Aseem Behl, Eshed Ohn-Bar, Kashyap Chitta, Andreas Geiger; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11763-11773

Data aggregation techniques can significantly improve vision-based policy learning within a training environment, e.g., learning to drive in a specific simulation condition. However, as on-policy data is sequentially sampled and added in an iterative manner, the policy can specialize and overfit to the training conditions. For real-world applications, it is useful for the learned policy to generalize to novel scenarios that differ from the training conditions. To improve policy learning while maintaining robustness when training end-to-end driving policies, we perform an extensive analysis of data aggregation techniques in the CARLA environment. We demonstrate how the majority of them have poor generalization performance, and develop a novel approach with empirically better generalization performance compared to existing techniques. Our two key ideas are (1) to sample critical states from the collected on-policy data based on the utility they provide to the learned policy in terms of driving behavior, and (2) to incorporate a replay buffer which progressively focuses on the high uncertainty regions of the policy's state distribution. We evaluate the proposed approach on the CARLA NoCrash benchmark, focusing on the most challenging driving scenarios with dense pedestrian and vehicle traffic. Our approach improves driving success rate by 16% over state-of-the-art, achieving 87% of the expert performance while also reducing the collision rate by an order of magnitude without the use of any additional modality, auxiliary tasks, architectural modifications or reward from the environment.

****

Geometrically Principled Connections in Graph Neural Networks

Shunwang Gong, Mehdi Bahri, Michael M. Bronstein, Stefanos Zafeiriou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11415-11424

Graph convolution operators bring the advantages of deep learning to a variety of graph and mesh processing tasks previously deemed out of reach. With their con

tinued success comes the desire to design more powerful architectures, often by adapting existing deep learning techniques to non-Euclidean data. In this paper, we argue geometry should remain the primary driving force behind innovation in the emerging field of geometric deep learning. We relate graph neural networks to widely successful computer graphics and data approximation models: radial basis functions (RBFs). We conjecture that, like RBFs, graph convolution layers would benefit from the addition of simple functions to the powerful convolution kernels. We introduce affine skip connections, a novel building block formed by combining a fully connected layer with any graph convolution operator. We experimentally demonstrate the effectiveness of our technique, and show the improved performance is the consequence of more than the increased number of parameters. Operators equipped with the affine skip connection markedly outperform their base performance on every task we evaluated, i.e., shape reconstruction, dense shape correspondence, and graph classification. We hope our simple and effective approach will serve as a solid baseline and help ease future research in graph neural networks.

****

Making Better Mistakes: Leveraging Class Hierarchies With Deep Networks

Luca Bertinetto, Romain Mueller, Konstantinos Tertikas, Sina Samangooei, Nicholas A. Lord; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12506-12515

Deep neural networks have improved image classification dramatically over the past decade, but have done so by focusing on performance measures that treat all classes other than the ground truth as equally wrong. This has led to a situation in which mistakes are less likely to be made than before, but are equally likely to be absurd or catastrophic when they do occur. Past works have recognised and tried to address this issue of mistake severity, often by using graph distances in class hierarchies, but this has largely been neglected since the advent of the current deep learning era in computer vision. In this paper, we aim to renew interest in this problem by reviewing past approaches and proposing two simple methods which outperform the prior art under several metrics on two large datasets with complex class hierarchies: tieredImageNet and iNaturalist'19.

****

Telling Left From Right: Learning Spatial Correspondence of Sight and Sound

Karren Yang, Bryan Russell, Justin Salamon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9932-9941

Self-supervised audio-visual learning aims to capture useful representations of video by leveraging correspondences between visual and audio inputs. Existing approaches have focused primarily on matching semantic information between the sensory streams. We propose a novel self-supervised task to leverage an orthogonal principle: matching spatial information in the audio stream to the positions of sound sources in the visual stream. Our approach is simple yet effective. We train a model to determine whether the left and right audio channels have been flipped, forcing it to reason about spatial localization across the visual and audio streams. To train and evaluate our method, we introduce a large-scale video dataset, YouTube-ASMR-300K, with spatial audio comprising over 900 hours of footage. We demonstrate that understanding spatial correspondence enables models to perform better on three audio-visual tasks, achieving quantitative gains over supervised and self-supervised baselines that do not leverage spatial audio cues. We also show how to extend our self-supervised approach to 360 degree videos with ambisonic audio.

****

Deep Adversarial Decomposition: A Unified Framework for Separating Superimposed Images

Zhengxia Zou, Sen Lei, Tianyang Shi, Zhenwei Shi, Jieping Ye; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12806-12816

Separating individual image layers from a single mixed image has long been an important but challenging task. We propose a unified framework named "deep adversarial decomposition" for single superimposed image separation. Our method deals w

ith both linear and non-linear mixtures under an adversarial training paradigm. Considering the layer separating ambiguity that given a single mixed input, ther e could be an infinite number of possible solutions, we introduce a "Separation-Critic" - a discriminative network which is trained to identify whether the outp ut layers are well-separated and thus further improves the layer separation. We also introduce a "crossroad L1" loss function, which computes the distance betwe en the unordered outputs and their references in a crossover manner so that the training can be well-instructed with pixel-wise supervision. Experimental result s suggest that our method significantly outperforms other popular image separati on frameworks. Without specific tuning, our method achieves the state of the art results on multiple computer vision tasks, including the image deraining, photo reflection removal, and image shadow removal.

****

Towards Accurate Scene Text Recognition With Semantic Reasoning Networks

Deli Yu, Xuan Li, Chengquan Zhang, Tao Liu, Junyu Han, Jingtuo Liu, Errui Ding; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Reco gnition (CVPR), 2020, pp. 12113-12122

Scene text image contains two levels of contents: visual texture and semantic in formation. Although the previous scene text recognition methods have made great progress over the past few years, the research on mining semantic information to assist text recognition attracts less attention, only RNN-like structures are e xplored to implicitly model semantic information. However, we observe that RNN b ased methods have some obvious shortcomings, such as time-dependent decoding man ner and one-way serial transmission of semantic context, which greatly limit the help of semantic information and the computation efficiency. To mitigate these limitations, we propose a novel end-to-end trainable framework named semantic re asoning network (SRN) for accurate scene text recognition, where a global semant ic reasoning module (GSRM) is introduced to capture global semantic context thro ugh multi-way parallel transmission. The state-of-the-art results on 7 public be nchmarks, including regular text, irregular text and non-Latin long text, verify the effectiveness and robustness of the proposed method. In addition, the speed of SRN has significant advantages over the RNN based methods, demonstrating its value in practical use.

****

Deep Relational Reasoning Graph Network for Arbitrary Shape Text Detection

Shi-Xue Zhang, Xiaobin Zhu, Jie-Bo Hou, Chang Liu, Chun Yang, Hongfa Wang, Xu-Cheng Yin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat tern Recognition (CVPR), 2020, pp. 9699-9708

Arbitrary shape text detection is a challenging task due to the high variety and complexity of scenes texts. In this paper, we propose a novel unified relationa l reasoning graph network for arbitrary shape text detection. In our method, an innovative local graph bridges a text proposal model via Convolutional Neural Ne twork (CNN) and a deep relational reasoning network via Graph Convolutional Netw ork (GCN), making our network end-to-end trainable. To be concrete, every text i nstance will be divided into a series of small rectangular components, and the g eometry attributes (e.g., height, width, and orientation) of the small component s will be estimated by our text proposal model. Given the geometry attributes, t he local graph construction model can roughly establish linkages between differe nt text components. For further reasoning and deducing the likelihood of linkage s between the component and its neighbors, we adopt a graph-based network to per form deep relational reasoning on local graphs. Experiments on public available datasets demonstrate the state-of-the-art performance of our method. Code is ava ilable at https://github.com/GXYM/DRRG.

****

GP-NAS: Gaussian Process Based Neural Architecture Search

Zhihang Li, Teng Xi, Jiankang Deng, Gang Zhang, Shengzhao Wen, Ran He; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2020, pp. 11933-11942

Neural architecture search (NAS) advances beyond the state-of-the-art in various computer vision tasks by automating the designs of deep neural networks. In thi

s paper, we aim to address three important questions in NAS: (1) How to measure the correlation between architectures and their performances? (2) How to evaluate the correlation between different architectures? (3) How to learn these correlations with a small number of samples? To this end, we first model these correlations from a Bayesian perspective. Specifically, by introducing a novel Gaussian Process based NAS (GP-NAS) method, the correlations are modeled by the kernel function and mean function. The kernel function is also learnable to enable adaptive modeling for complex correlations in different search spaces. Furthermore, by incorporating a mutual information based sampling method, we can theoretically ensure the high-performance architecture with only a small set of samples. After addressing these problems, training GP-NAS once enables direct performance prediction of any architecture in different scenarios and may obtain efficient networks for different deployment platforms. Extensive experiments on both image classification and face recognition tasks verify the effectiveness of our algorithm.

****

Basis Prediction Networks for Effective Burst Denoising With Large Kernels
Zhihao Xia, Federico Perazzi, Michael Gharbi, Kalyan Sunkavalli, Ayan Chakrabarti; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11844-11853

Bursts of images exhibit significant self-similarity across both time and space. This motivates a representation of the kernels as linear combinations of a small set of basis elements. To this end, we introduce a novel basis prediction network that, given an input burst, predicts a set of global basis kernels --- shared within the image --- and the corresponding mixing coefficients --- which are specific to individual pixels. Compared to state-of-the-art techniques that output a large tensor of per-pixel spatiotemporal kernels, our formulation substantially reduces the dimensionality of the network output. This allows us to effectively exploit comparatively larger denoising kernels, achieving both significant quality improvements (over 1dB PSNR) and faster run-times over state-of-the-art methods.

****

Real-World Person Re-Identification via Degradation Invariance Learning
Yukun Huang, Zheng-Jun Zha, Xueyang Fu, Richang Hong, Liang Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14084-14094

Person re-identification (Re-ID) in real-world scenarios usually suffers from various degradation factors, e.g., low-resolution, weak illumination, blurring and adverse weather. On the one hand, these degradations lead to severe discriminative information loss, which significantly obstructs identity representation learning; on the other hand, the feature mismatch problem caused by low-level visual variations greatly reduces retrieval performance. An intuitive solution to this problem is to utilize low-level image restoration methods to improve the image quality. However, existing restoration methods cannot directly serve to real-world Re-ID due to various limitations, e.g., the requirements of reference samples, domain gap between synthesis and reality, and incompatibility between low-level and high-level methods. In this paper, to solve the above problem, we propose a degradation invariance learning framework for real-world person Re-ID. By introducing a self-supervised disentangled representation learning strategy, our method is able to simultaneously extract identity-related robust features and remove real-world degradations without extra supervision. We use low-resolution images as the main demonstration, and experiments show that our approach is able to achieve state-of-the-art performance on several Re-ID benchmarks. In addition, our framework can be easily extended to other real-world degradation factors, such as weak illumination, with only a few modifications.

****

Momentum Contrast for Unsupervised Visual Representation Learning
Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, Ross Girshick; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9729-9738

We present Momentum Contrast (MoCo) for unsupervised visual representation learn ing. From a perspective on contrastive learning as dictionary look-up, we build a dynamic dictionary with a queue and a moving-averaged encoder. This enables bu ilding a large and consistent dictionary on-the-fly that facilitates contrastive unsupervised learning. MoCo provides competitive results under the common linea r protocol on ImageNet classification. More importantly, the representations lea rned by MoCo transfer well to downstream tasks. MoCo can outperform its supervis ed pre-training counterpart in 7 detection/segmentation tasks on PASCAL VOC, COC O, and other datasets, sometimes surpassing it by large margins. This suggests t hat the gap between unsupervised and supervised representation learning has been largely closed in many vision tasks.
****

Meta-Learning of Neural Architectures for Few-Shot Learning
Thomas Elsken, Benedikt Staffler, Jan Hendrik Metzen, Frank Hutter; Proceedin gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12365-12375
The recent progress in neural architecture search (NAS) has allowed scaling the automated design of neural architectures to real-world domains, such as object d etection and semantic segmentation. However, one prerequisite for the applicatio n of NAS are large amounts of labeled data and compute resources. This renders i ts application challenging in few-shot learning scenarios, where many related ta sks need to be learned, each with limited amounts of data and compute time. Thus , few-shot learning is typically done with a fixed neural architecture. To impro ve upon this, we propose MetaNAS, the first method which fully integrates NAS wi th gradient-based meta-learning. MetaNAS optimizes a meta-architecture along wit h the meta-weights during meta-training. During meta-testing, architectures can be adapted to a novel task with a few steps of the task optimizer, that is: task adaptation becomes computationally cheap and requires only little data per task . Moreover, MetaNAS is agnostic in that it can be used with arbitrary model-agno stic meta-learning algorithms and arbitrary gradient-based NAS methods. Empirica l results on standard few-shot classification benchmarks show that MetaNAS with a combination of DARTS and REPTILE yields state-of-the-art results.
****

Deep Generative Model for Robust Imbalance Classification
Xinyue Wang, Yilin Lyu, Liping Jing; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14124-14133
Discovering hidden pattern from imbalanced data is a critical issue in various r eal-world applications including computer vision. The existing classification me thods usually suffer from the limitation of data especially the minority classes , and result in unstable prediction and low performance. In this paper, a deep g enerative classifier is proposed to mitigate this issue via both data perturbati on and model perturbation. Specially, the proposed generative classifier is mode led by a deep latent variable model where the latent variable aims to capture th e direct cause of target label. Meanwhile, the latent variable is represented by a probability distribution over possible values rather than a single fixed valu e, which is able to enforce uncertainty of model and lead to stable prediction. Furthermore, this latent variable, as a confounder, affects the process of data (feature/label) generation, so that we can arrive at well-justified sampling var iability considerations in statistics, and implement data perturbation. Extensiv e experiments have been conducted on widely-used real imbalanced image datasets. By comparing with the state-of-the-art methods, experimental results demonstrat e the superiority of our proposed model on imbalance classification task.
****

Unsupervised Multi-Modal Image Registration via Geometry Preserving Image-to-Ima ge Translation
Moab Arar, Yiftach Ginger, Dov Danon, Amit H. Bermano, Daniel Cohen-Or; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2020, pp. 13410-13419
Many applications, such as autonomous driving, heavily rely on multi-modal data where spatial alignment between the modalities is required. Most multi-modal reg

istration methods struggle computing the spatial correspondence between the imag es using prevalent cross-modality similarity measures. In this work, we bypass t he difficulties of developing cross-modality similarity measures, by training an image-to-image translation network on the two input modalities. This learned tr anslation allows training the registration network using simple and reliable mon o-modality metrics. We perform multi-modal registration using two networks - a s patial transformation network and a translation network. We show that by encoura ging our translation network to be geometry preserving, we manage to train an ac curate spatial transformation network. Compared to state-of-the-art multi-modal methods our presented method is unsupervised, requiring no pairs of aligned moda lities for training, and can be adapted to any pair of modalities. We evaluate o ur method quantitatively and qualitatively on commercial datasets, showing that it performs well on several modalities and achieves accurate alignment.
****

SCATTER: Selective Context Attentional Scene Text Recognizer
Ron Litman, Oron Anschel, Shahar Tsiper, Roee Litman, Shai Mazor, R. Manmat ha; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn ition (CVPR), 2020, pp. 11962-11972
Scene Text Recognition (STR), the task of recognizing text against complex image backgrounds, is an active area of research. Current state-of-the-art (SOTA) met hods still struggle to recognize text written in arbitrary shapes. In this paper , we introduce a novel architecture for STR, named Selective Context ATtentional Text Recognizer (SCATTER). SCATTER utilizes a stacked block architecture with i ntermediate supervision during training, that paves the way to successfully trai n a deep BiLSTM encoder, thus improving the encoding of contextual dependencies. Decoding is done using a two-step 1D attention mechanism. The first attention s tep re-weights visual features from a CNN backbone together with contextual feat ures computed by a BiLSTM layer. The second attention step, similar to previous papers, treats the features as a sequence and attends to the intra-sequence rela tionships. Experiments show that the proposed approach surpasses SOTA performanc e on irregular text recognition benchmarks by 3.7% on average.
****

Incremental Few-Shot Object Detection
Juan-Manuel Perez-Rua, Xiatian Zhu, Timothy M. Hospedales, Tao Xiang; Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13846-13855
Existing object detection methods typically rely on the availability of abundant labelled training samples per class and offline model training in a batch mode. These requirements substantially limit their scalability to open-ended accommod ation of novel classes with limited labelled training data, both in terms of mod el accuracy and training efficiency during deployment. We present the first stud y aiming to go beyond these limitations by considering the Incremental Few-Shot Detection (iFSD) problem setting, where new classes must be registered increment ally (without revisiting base classes) and with few examples. To this end we pro pose OpeN-ended Centre nEt (ONCE), a detector designed for incrementally learnin g to detect novel class objects with few examples. This is achieved by an elegan t adaptation of the efficient CentreNet detector to the few-shot learning scenar io, and meta-learning a class-wise code generator model for registering novel cl asses. ONCE fully respects the incremental learning paradigm, with novel class r egistration requiring only a single forward pass of few-shot training samples, a nd no access to base classes - thus making it suitable for deployment on embedde d devices, etc. Extensive experiments conducted on both the standard object dete ction (COCO, PASCAL VOC) and fashion landmark detection (DeepFashion2) tasks sho w the feasibility of iFSD for the first time, opening an interesting and very im portant line of research.
****

Cloth in the Wind: A Case Study of Physical Measurement Through Simulation
Tom F. H. Runia, Kirill Gavrilyuk, Cees G. M. Snoek, Arnold W. M. Smeulders; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio n (CVPR), 2020, pp. 10498-10507

For many of the physical phenomena around us, we have developed sophisticated models explaining their behavior. Nevertheless, measuring physical properties from visual observations is challenging due to the high number of causally underlying physical parameters -- including material properties and external forces. In this paper, we propose to measure latent physical properties for cloth in the wind without ever having seen a real example before. Our solution is an iterative refinement procedure with simulation at its core. The algorithm gradually updates the physical model parameters by running a simulation of the observed phenomenon and comparing the current simulation to a real-world observation. The correspondence is measured using an embedding function that maps physically similar examples to nearby points. We consider a case study of cloth in the wind, with curling flags as our leading example -- a seemingly simple phenomena but physically highly involved. Based on the physics of cloth and its visual manifestation, we propose an instantiation of the embedding function. For this mapping, modeled as a deep network, we introduce a spectral layer that decomposes a video volume into its temporal spectral power and corresponding frequencies. Our experiments demonstrate that the proposed method compares favorably to prior work on the task of measuring cloth material properties and external wind force from a real-world video.

****

Generalized Zero-Shot Learning via Over-Complete Distribution

Rohit Keshari, Richa Singh, Mayank Vatsa; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13300-13308

A well trained and generalized deep neural network (DNN) should be robust to both seen and unseen classes. However, the performance of most of the existing supervised DNN algorithms degrade for classes which are unseen in the training set. To learn a discriminative classifier which yields good performance in Zero-Shot Learning (ZSL) settings, we propose to generate an Over-Complete Distribution (OCD) using Conditional Variational Autoencoder (CVAE) of both seen and unseen classes. In order to enforce the separability between classes and reduce the class scatter, we propose the use of Online Batch Triplet Loss (OBTL) and Center Loss (CL) on the generated OCD. The effectiveness of the framework is evaluated using both Zero-Shot Learning and Generalized Zero-Shot Learning protocols on three publicly available benchmark databases, SUN, CUB and AWA2. The results show that generating over-complete distributions and enforcing the classifier to learn a transform function from overlapping to non-overlapping distributions can improve the performance on both seen and unseen classes.

****

On the General Value of Evidence, and Bilingual Scene-Text Visual Question Answering

Xinyu Wang, Yuliang Liu, Chunhua Shen, Chun Chet Ng, Canjie Luo, Lianwen Jin, Chee Seng Chan, Anton van den Hengel, Liangwei Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10126-10135

Visual Question Answering (VQA) methods have made incredible progress, but suffer from a failure to generalize. This is visible in the fact that they are vulnerable to learning coincidental correlations in the data rather than deeper relations between image content and ideas expressed in language. We present a dataset that takes a step towards addressing this problem in that it contains questions expressed in two languages, and an evaluation process that co-opts a well understood image-based metric to reflect the method's ability to reason. Measuring reasoning directly encourages generalization by penalizing answers that are coincidentally correct. The dataset reflects the scene-text version of the VQA problem, and the reasoning evaluation can be seen as a text-based version of a referring expression challenge. Experiments and analyses are provided that show the value of the dataset. The dataset is available at www.est-vqa.org.

****

Designing Network Design Spaces

Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, Piotr Dollar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Reco

gnition (CVPR), 2020, pp. 10428-10436
In this work, we present a new network design paradigm. Our goal is to help advance the understanding of network design and discover design principles that generalize across settings. Instead of focusing on designing individual network instances, we design network design spaces that parametrize populations of networks. The overall process is analogous to classic manual design of networks, but elevated to the design space level. Using our methodology we explore the structure aspect of network design and arrive at a low-dimensional design space consisting of simple, regular networks that we call RegNet. The core insight of the RegNet parametrization is surprisingly simple: widths and depths of good networks can be explained by a quantized linear function. We analyze the RegNet design space and arrive at interesting findings that do not match the current practice of network design. The RegNet design space provides simple and fast networks that work well across a wide range of flop regimes. Under comparable training settings and flops, the RegNet models outperform the popular EfficientNet models while being up to 5x faster on GPUs.
****

Regularizing CNN Transfer Learning With Randomised Regression
Yang Zhong, Atsuto Maki; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13637-13646
This paper is about regularizing deep convolutional networks (CNNs) based on an adaptive framework for transfer learning with limited training data in the target domain. Recent advances of CNN regularization in this context are commonly due to the use of additional regularization objectives. They guide the training away from the target task using some forms of concrete tasks. Unlike those related approaches, we suggest that an objective without a concrete goal can still serve well as a regularizer. In particular, we demonstrate Pseudo-task Regularization (PtR) which dynamically regularizes a network by simply attempting to regress image representations to pseudo-regression targets during fine-tuning. That is, a CNN is efficiently regularized without additional resources of data or prior domain expertise. In sum, the proposed PtR provides: a) an alternative for network regularization without dependence on the design of concrete regularization objectives or extra annotations; b) a dynamically adjusted and maintained strength of regularization effect by balancing the gradient norms between objectives on-line. Through numerous experiments, surprisingly, the improvements on classification accuracy by PtR are shown greater or on a par to the recent state-of-the-art methods.
****

PVN3D: A Deep Point-Wise 3D Keypoints Voting Network for 6DoF Pose Estimation
Yisheng He, Wei Sun, Haibin Huang, Jianran Liu, Haoqiang Fan, Jian Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11632-11641
In this work, we present a novel data-driven method for robust 6DoF object pose estimation from a single RGBD image. Unlike previous methods that directly regressing pose parameters, we tackle this challenging task with a keypoint-based approach. Specifically, we propose a deep Hough voting network to detect 3D keypoints of objects and then estimate the 6D pose parameters within a least-squares fitting manner. Our method is a natural extension of 2D-keypoint approaches that successfully work on RGB based 6DoF estimation. It allows us to fully utilize the geometric constraint of rigid objects with the extra depth information and is easy for a network to learn and optimize. Extensive experiments were conducted to demonstrate the effectiveness of 3D-keypoint detection in the 6D pose estimation task. Experimental results also show our method outperforms the state-of-the-art methods by large margins on several benchmarks. Code and video are available at https://github.com/ethnhe/PVN3D.git.
****

Domain-Aware Visual Bias Eliminating for Generalized Zero-Shot Learning
Shaobo Min, Hantao Yao, Hongtao Xie, Chaoqun Wang, Zheng-Jun Zha, Yongdong Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12664-12673

Generalized zero-shot learning aims to recognize images from seen and unseen dom ains. Recent methods focus on learning a unified semantic-aligned visual represe ntation to transfer knowledge between two domains, while ignoring the effect of semantic-free visual representation in alleviating the biased recognition proble m. In this paper, we propose a novel Domain-aware Visual Bias Eliminating (DVBE) network that constructs two complementary visual representations, i.e., semanti c-free and semantic-aligned, to treat seen and unseen domains separately. Specif ically, we explore cross-attentive second-order visual statistics to compact the semantic-free representation, and design an adaptive margin Softmax to maximize inter-class divergences. Thus, the semantic-free representation becomes discrim inative enough to not only predict seen class accurately but also filter out uns een images, i.e., domain detection, based on the predicted class entropy. For un seen images, we automatically search an optimal semantic-visual alignment archit ecture, rather than manual designs, to predict unseen classes. With accurate dom ain detection, the biased recognition problem towards the seen domain is signifi cantly reduced. Experiments on five benchmarks for classification and segmentati on show that DVBE outperforms existing methods by averaged 5.7% improvement.

****

VSGNet: Spatial Attention Network for Detecting Human Object Interactions Using Graph Convolutions

Oytun Ulutan, A S M Iftekhar, B. S. Manjunath; Proceedings of the IEEE/CVF Con ference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13617-13626

Comprehensive visual understanding requires detection frameworks that can effect ively learn and utilize object interactions while analyzing objects individually . This is the main objective in Human-Object Interaction (HOI) detection task. I n particular, relative spatial reasoning and structural connections between obje cts are essential cues for analyzing interactions, which is addressed by the pro posed Visual-Spatial-Graph Network (VSGNet) architecture. VSGNet extracts visual features from the human-object pairs, refines the features with spatial configu rations of the pair, and utilizes the structural connections between the pair vi a graph convolutions. The performance of VSGNet is thoroughly evaluated using th e Verbs in COCO (V-COCO) dataset. Experimental results indicate that VSGNet outp erforms state-of-the-art solutions by 8% or 4 mAP.

****

Few-Shot Video Classification via Temporal Alignment

Kaidi Cao, Jingwei Ji, Zhangjie Cao, Chien-Yi Chang, Juan Carlos Niebles; Pr oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10618-10627

Difficulty in collecting and annotating large-scale video data raises a growing interest in learning models which can recognize novel classes with only a few tr aining examples. In this paper, we propose the Ordered Temporal Alignment Module (OTAM), a novel few-shot learning framework that can learn to classify a previo usly unseen video. While most previous work neglects long-term temporal ordering information, our proposed model explicitly leverages the temporal ordering info rmation in video data through ordered temporal alignment. This leads to strong d ata-efficiency for few-shot learning. In concrete, our proposed pipeline learns a deep distance measurement of the query video with respect to novel class proxi es over its alignment path. We adopt an episode-based training scheme and direct ly optimize the few-shot learning objective. We evaluate OTAM on two challenging real-world datasets, Kinetics and Something-Something-V2, and show that our mod el leads to significant improvement of few-shot video classification over a wide range of competitive baselines and outperforms state-of-the-art benchmarks by a large margin.

****

Density-Aware Graph for Deep Semi-Supervised Visual Recognition

Suichan Li, Bin Liu, Dongdong Chen, Qi Chu, Lu Yuan, Nenghai Yu; Proceeding s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13400-13409

Semi-supervised learning (SSL) has been extensively studied to improve the gener alization ability of deep neural networks for visual recognition. To involve the

unlabelled data, most existing SSL methods are based on common density-based cluster assumption: samples lying in the same high-density region are likely to belong to the same class, including the methods performing consistency regularization or generating pseudo-labels for the unlabelled images. Despite their impressive performance, we argue three limitations exist: 1) Though the density information is demonstrated to be an important clue, they all use it in an implicit way and have not exploited it in depth. 2) For feature learning, they often learn the feature embedding based on the single data sample and ignore the neighborhood information. 3) For label-propagation based pseudo-label generation, it is often done offline and difficult to be end-to-end trained with feature learning. Motivated by these limitations, this paper proposes to solve the SSL problem by building a novel density-aware graph, based on which the neighborhood information can be easily leveraged and the feature learning and label propagation can also be trained in an end-to-end way. Specifically, we first propose a new Density-aware Neighborhood Aggregation(DNA) module to learn more discriminative features by incorporating the neighborhood information in a density-aware manner. Then a novel Density-ascending Path based Label Propagation(DPLP) module is proposed to generate the pseudo-labels for unlabeled samples more efficiently according to the feature distribution characterized by density. Finally, the DNA module and DPLP module evolve and improve each other end-to-end. Extensive experiments demonstrate the effectiveness of the newly proposed density-aware graph based SSL framework and our approach can outperform current state-of-the-art methods by a large margin.

****

Learning Deep Network for Detecting 3D Object Keypoints and 6D Poses
Wanqing Zhao, Shaobo Zhang, Ziyu Guan, Wei Zhao, Jinye Peng, Jianping Fan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14134-14142

The state-of-art 6D object pose detection methods use convolutional neural networks to estimate objects' 6D poses from RGB images. However, they require huge numbers of images with explicit 3D annotations such as 6D poses, 3D bounding boxes and 3D keypoints, either obtained by manual labeling or inferred from synthetic images generated by 3D CAD models. Manual labeling for a large number of images is a laborious task, and we usually do not have the corresponding 3D CAD models of objects in real environment. In this paper, we develop a keypoint-based 6D object pose detection method (and its deep network) called Object Keypoint based POSe Estimation (OK-POSE). OK-POSE employs relative transformation between viewpoints for training. Specifically, we use pairs of images with object annotation and relative transformation information between their viewpoints to automatically discover objects' 3D keypoints which are geometrically and visually consistent. Then, the 6D object pose can be estimated using a keypoint-based geometric reasoning method with a reference viewpoint. The relative transformation information can be easily obtained from any cheap binocular cameras or most smartphone devices, thus greatly lowering the labeling cost. Experiments have demonstrated that OK-POSE achieves acceptable performance compared to methods relying on the object's 3D CAD model or a great deal of 3D labeling. These results show that our method can be used as a suitable alternative when there are no 3D CAD models or a large number of 3D annotations.

****

REVERIE: Remote Embodied Visual Referring Expression in Real Indoor Environments
Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, Anton van den Hengel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9982-9991

One of the long-term challenges of robotics is to enable robots to interact with humans in the visual world via natural language, as humans are visual animals that communicate through language. Overcoming this challenge requires the ability to perform a wide variety of complex tasks in response to multifarious instructions from humans. In the hope that it might drive progress towards more flexible and powerful human interactions with robots, we propose a dataset of varied and complex robot tasks, described in natural language, in terms of objects visible

in a large set of real images. Given an instruction, success requires navigating through a previously-unseen environment to identify an object. This represents a practical challenge, but one that closely reflects one of the core visual problems in robotics. Several state-of-the-art vision-and-language navigation, and referring-expression models are tested to verify the difficulty of this new task, but none of them show promising results because there are many fundamental differences between our task and previous ones. A novel Interactive Navigator-Pointer model is also proposed that provides a strong baseline on the task. The proposed model especially achieves the best performance on the unseen test split, but still leaves substantial room for improvement compared to the human performance. Repository: https://github.com/YuankaiQi/REVERIE.

****

Deep Iterative Surface Normal Estimation

Jan Eric Lenssen, Christian Osendorfer, Jonathan Masci; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11247-11256

This paper presents an end-to-end differentiable algorithm for robust and detail-preserving surface normal estimation on unstructured point-clouds. We utilize graph neural networks to iteratively parameterize an adaptive anisotropic kernel that produces point weights for weighted least-squares plane fitting in local neighborhoods. The approach retains the interpretability and efficiency of traditional sequential plane fitting while benefiting from adaptation to data set statistics through deep learning. This results in a state-of-the-art surface normal estimator that is robust to noise, outliers and point density variation, preserves sharp features through anisotropic kernels and equivariance through a local quaternion-based spatial transformer. Contrary to previous deep learning methods, the proposed approach does not require any hand-crafted features or preprocessing. It improves on the state-of-the-art results while being more than two orders of magnitude faster and more parameter efficient.

****

Unified Dynamic Convolutional Network for Super-Resolution With Variational Degradations

Yu-Syuan Xu, Shou-Yao Roy Tseng, Yu Tseng, Hsien-Kai Kuo, Yi-Min Tsai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12496-12505

Deep Convolutional Neural Networks (CNNs) have achieved remarkable results on Single Image Super-Resolution (SISR). Despite considering only a single degradation, recent studies also include multiple degrading effects to better reflect real-world cases. However, most of the works assume a fixed combination of degrading effects, or even train an individual network for different combinations. Instead, a more practical approach is to train a single network for wide-ranging and variational degradations. To fulfill this requirement, this paper proposes a unified network to accommodate the variations from inter-image (cross-image variations) and intra-image (spatial variations). Different from the existing works, we incorporate dynamic convolution which is a far more flexible alternative to handle different variations. In SISR with non-blind setting, our Unified Dynamic Convolutional Network for Variational Degradations (UDVD) is evaluated on both synthetic and real images with an extensive set of variations. The qualitative results demonstrate the effectiveness of UDVD over various existing works. Extensive experiments show that our UDVD achieves favorable or comparable performance on both synthetic and real images.

****

Noisier2Noise: Learning to Denoise From Unpaired Noisy Data

Nick Moran, Dan Schmidt, Yu Zhong, Patrick Coady; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12064-12072

We present a method for training a neural network to perform image denoising without access to clean training examples or access to paired noisy training examples. Our method requires only a single noisy realization of each training example and a statistical model of the noise distribution, and is applicable to a wide

variety of noise models, including spatially structured noise. Our model produces results which are competitive with other learned methods which require richer training data, and outperforms traditional non-learned denoising methods. We present derivations of our method for arbitrary additive noise, an improvement specific to Gaussian additive noise, and an extension to multiplicative Bernoulli noise.

****

PhysGAN: Generating Physical-World-Resilient Adversarial Examples for Autonomous Driving

Zelun Kong, Junfeng Guo, Ang Li, Cong Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14254-14263

Although Deep neural networks (DNNs) are being pervasively used in vision-based autonomous driving systems, they are found vulnerable to adversarial attacks where small-magnitude perturbations into the inputs during test time cause dramatic changes to the outputs. While most of the recent attack methods target at digital-world adversarial scenarios, it is unclear how they perform in the physical world, and more importantly, the generated perturbations under such methods would cover a whole driving scene including those fixed background imagery such as the sky, making them inapplicable to physical world implementation. We present PhysGAN, which generates physical-world-resilient adversarial examples for misleading autonomous driving systems in a continuous manner. We show the effectiveness and robustness of PhysGAN via extensive digital- and real-world evaluations. We compare PhysGAN with a set of state-of-the-art baseline methods, which further demonstrate the robustness and efficacy of our approach. We also show that PhysGAN outperforms state-of-the-art baseline methods. To the best of our knowledge, PhysGAN is probably the first technique of generating realistic and physical-world-resilient adversarial examples for attacking common autonomous driving scenarios.

****

Fast(er) Reconstruction of Shredded Text Documents via Self-Supervised Deep Asymmetric Metric Learning

Thiago M. Paixao, Rodrigo F. Berriel, Maria C. S. Boeres, Alessandro L. Koerich, Claudine Badue, Alberto F. De Souza, Thiago Oliveira-Santos; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14343-14351

The reconstruction of shredded documents consists in arranging the pieces of paper (shreds) in order to reassemble the original aspect of such documents. This task is particularly relevant for supporting forensic investigation as documents may contain criminal evidence. As an alternative to the laborious and time-consuming manual process, several researchers have been investigating ways to perform automatic digital reconstruction. A central problem in automatic reconstruction of shredded documents is the pairwise compatibility evaluation of the shreds, notably for binary text documents. In this context, deep learning has enabled great progress for accurate reconstructions in the domain of mechanically-shredded documents. A sensitive issue, however, is that current deep model solutions require an inference whenever a pair of shreds has to be evaluated. This work proposes a scalable deep learning approach for measuring pairwise compatibility in which the number of inferences scales linearly (rather than quadratically) with the number of shreds. Instead of predicting compatibility directly, deep models are leveraged to asymmetrically project the raw shred content onto a common metric space in which distance is proportional to the compatibility. Experimental results show that our method has accuracy comparable to the state-of-the-art with a speed-up of about 22 times for a test instance with 505 shreds (20 mixed shredded-pages from different documents).

****

MoreFusion: Multi-object Reasoning for 6D Pose Estimation from Volumetric Fusion

Kentaro Wada, Edgar Sucar, Stephen James, Daniel Lenton, Andrew J. Davison; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14540-14549

Robots and other smart devices need efficient object-based scene representations

from their on-board vision systems to reason about contact, physics and occlusion. Recognized precise object models will play an important role alongside non-parametric reconstructions of unrecognized structures. We present a system which can estimate the accurate poses of multiple known objects in contact and occlusion from real-time, embodied multi-view vision. Our approach makes 3D object pose proposals from single RGB-D views, accumulates pose estimates and non-parametric occupancy information from multiple views as the camera moves, and performs joint optimization to estimate consistent, non-intersecting poses for multiple objects in contact. We verify the accuracy and robustness of our approach experimentally on 2 object datasets: YCB-Video, and our own challenging Cluttered YCB-Video. We demonstrate a real-time robotics application where a robot arm precisely and orderly disassembles complicated piles of objects, using only on-board RGB-D vision.

****

Filter Response Normalization Layer: Eliminating Batch Dependence in the Training of Deep Neural Networks

Saurabh Singh, Shankar Krishnan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11237-11246

Batch Normalization (BN) uses mini-batch statistics to normalize the activations during training, introducing dependence between mini-batch elements. This dependency can hurt the performance if the mini-batch size is too small, or if the elements are correlated. Several alternatives, such as Batch Renormalization and Group Normalization (GN), have been proposed to address this issue. However, they either do not match the performance of BN for large batches, or still exhibit degradation in performance for smaller batches, or introduce artificial constraints on the model architecture. In this paper we propose the Filter Response Normalization (FRN) layer, a novel combination of a normalization and an activation function, that can be used as a replacement for other normalizations and activations. Our method operates on each activation channel of each batch element independently, eliminating the dependency on other batch elements. Our method outperforms BN and other alternatives in a variety of settings for all batch sizes. FRN layer performs 0.7-1.0% better than BN on top-1 validation accuracy with large mini-batch sizes for Imagenet classification using InceptionV3 and ResnetV2-50 architectures. Further, it performs >1% better than GN on the same problem in the small mini-batch size regime. For object detection problem on COCO dataset, FRN layer outperforms all other methods by at least 0.3-0.5% in all batch size regimes.

****

Visual Reaction: Learning to Play Catch With Your Drone

Kuo-Hao Zeng, Roozbeh Mottaghi, Luca Weihs, Ali Farhadi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11573-11582

In this paper we address the problem of visual reaction: the task of interacting with dynamic environments where the changes in the environment are not necessarily caused by the agents itself. Visual reaction entails predicting the future changes in a visual environment and planning accordingly. We study the problem of visual reaction in the context of playing catch with a drone in visually rich synthetic environments. This is a challenging problem since the agent is required to learn (1) how objects with different physical properties and shapes move, (2) what sequence of actions should be taken according to the prediction, (3) how to adjust the actions based on the visual feedback from the dynamic environment (e.g., when objects bouncing off a wall), and (4) how to reason and act with an unexpected state change in a timely manner. We propose a new dataset for this task, which includes 30K throws of 20 types of objects in different directions with different forces. Our results show that our model that integrates a forecaster with a planner outperforms a set of strong baselines that are based on tracking as well as pure model-based and model-free RL baselines. The code and dataset are available at github.com/KuoHaoZeng/Visual_Reaction.

****

Learning to See Through Obstructions

Yu-Lun Liu, Wei-Sheng Lai, Ming-Hsuan Yang, Yung-Yu Chuang, Jia-Bin Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14215-14224

We present a learning-based approach for removing unwanted obstructions, such as window reflections, fence occlusions or raindrops, from a short sequence of images captured by a moving camera. Our method leverages the motion differences between the background and the obstructing elements to recover both layers. Specifically, we alternate between estimating dense optical flow fields of the two layers and reconstructing each layer from the flow-warped images via a deep convolutional neural network. The learning-based layer reconstruction allows us to accommodate potential errors in the flow estimation and brittle assumptions such as brightness consistency. We show that training on synthetically generated data transfers well to real images. Our results on numerous challenging scenarios of reflection and fence removal demonstrate the effectiveness of the proposed method.

****

SpeedNet: Learning the Speediness in Videos

Sagie Benaim, Ariel Ephrat, Oran Lang, Inbar Mosseri, William T. Freeman, Michael Rubinstein, Michal Irani, Tali Dekel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9922-9931

We wish to automatically predict the "speediness" of moving objects in videos - whether they move faster, at, or slower than their "natural" speed. The core component in our approach is SpeedNet--a novel deep network trained to detect if a video is playing at normal rate, or if it is sped up. SpeedNet is trained on a large corpus of natural videos in a self-supervised manner, without requiring any manual annotations. We show how this single, binary classification network can be used to detect arbitrary rates of speediness of objects. We demonstrate prediction results by SpeedNet on a wide range of videos containing complex natural motions, and examine the visual cues it utilizes for making those predictions. Importantly, we show that through predicting the speed of videos, the model learns a powerful and meaningful space-time representation that goes beyond simple motion cues. We demonstrate how those learned features can boost the performance of self supervised action recognition, and can be used for video retrieval. Furthermore, we also apply SpeedNet for generating time-varying, adaptive video speedups, which can allow viewers to watch videos faster, but with less of the jittery, unnatural motions typical to videos that are sped up uniformly.

****

IMRAM: Iterative Matching With Recurrent Attention Memory for Cross-Modal Image-Text Retrieval

Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, Jungong Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12655-12663

Enabling bi-directional retrieval of images and texts is important for understanding the correspondence between vision and language. Existing methods leverage the attention mechanism to explore such correspondence in a fine-grained manner. However, most of them consider all semantics equally and thus align them uniformly, regardless of their diverse complexities. In fact, semantics are diverse (i.e. involving different kinds of semantic concepts), and humans usually follow a latent structure to combine them into understandable languages. It may be difficult to optimally capture such sophisticated correspondences in existing methods. In this paper, to address such a deficiency, we propose an Iterative Matching with Recurrent Attention Memory (IMRAM) method, in which correspondences between images and texts are captured with multiple steps of alignments. Specifically, we introduce an iterative matching scheme to explore such fine-grained correspondence progressively. A memory distillation unit is used to refine alignment knowledge from early steps to later ones. Experiment results on three benchmark datasets, i.e. Flickr8K, Flickr30K, and MS COCO, show that our IMRAM achieves state-of-the-art performance, well demonstrating its effectiveness. Experiments on a practical business advertisement dataset, named KWAI-AD, further validates the applicability of our method in practical scenarios.

****

Satellite Image Time Series Classification With Pixel-Set Encoders and Temporal Self-Attention
Vivien Sainte Fare Garnot, Loic Landrieu, Sebastien Giordano, Nesrine Chehata ; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12325-12334
Satellite image time series, bolstered by their growing availability, are at the forefront of an extensive effort towards automated Earth monitoring by international institutions. In particular, large-scale control of agricultural parcels is an issue of major political and economic importance. In this regard, hybrid convolutional-recurrent neural architectures have shown promising results for the automated classification of satellite image time series. We propose an alternative approach in which the convolutional layers are advantageously replaced with encoders operating on unordered sets of pixels to exploit the typically coarse resolution of publicly available satellite images. We also propose to extract temporal features using a bespoke neural architecture based on self-attention instead of recurrent networks. We demonstrate experimentally that our method not only outperforms previous state-of-the-art approaches in terms of precision, but also significantly decreases processing time and memory requirements. Lastly, we release a large open-access annotated dataset as a benchmark for future work on satellite image time series.
****
Train in Germany, Test in the USA: Making 3D Object Detectors Generalize
Yan Wang, Xiangyu Chen, Yurong You, Li Erran Li, Bharath Hariharan, Mark Campbell, Kilian Q. Weinberger, Wei-Lun Chao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11713-11723
In the domain of autonomous driving, deep learning has substantially improved the 3D object detection accuracy for LiDAR and stereo camera data alike. While deep networks are great at generalization, they are also notorious to overfit to all kinds of spurious artifacts, such as brightness, car sizes and models, that may appear consistently throughout the data. In fact, most datasets for autonomous driving are collected within a narrow subset of cities within one country, typically under similar weather conditions. In this paper we consider the task of adapting 3D object detectors from one dataset to another. We observe that naively, this appears to be a very challenging task, resulting in drastic drops in accuracy levels. We provide extensive experiments to investigate the true adaptation challenges and arrive at a surprising conclusion: the primary adaptation hurdle to overcome are differences in car sizes across geographic areas. A simple correction based on the average car size yields a strong correction of the adaptation gap. Our proposed method is simple and easily incorporated into most 3D object detection frameworks. It provides a first baseline for 3D object detection adaptation across countries, and gives hope that the underlying problem may be more within grasp than one may have hoped to believe. Our code is available at https://github. com/cxy1997/3D_adapt_auto_driving.
****
CARP: Compression Through Adaptive Recursive Partitioning for Multi-Dimensional Images
Rongjie Liu, Meng Li, Li Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14306-14314
Fast and effective image compression for multi-dimensional images has become increasingly important for efficient storage and transfer of massive amounts of high resolution images and videos. Desirable properties in compression methods include (1) high reconstruction quality at a wide range of compression rates while preserving key local details, (2) computational scalability, (3) applicability to a variety of different image/video types and of different dimensions, and (4) ease of tuning. We present such a method for multi-dimensional image compression called Compression via Adaptive Recursive Partitioning (CARP). CARP uses an optimal permutation of the image pixels inferred from a Bayesian probabilistic model on recursive partitions of the image to reduce its effective dimensionality, achieving a parsimonious representation that preserves information. CARP uses a multi-layer Bayesian hierarchical model to achieve self-tuning and regularization

to avoid overfitting-- resulting in one single parameter to be specified by the user to achieve the desired compression rate. Extensive numerical experiments using a variety of datasets including 2D ImageNet, 3D medical image, and real-life YouTube and surveillance videos show that CARP dominates the state-of-the-art compression approaches-- including JPEG, JPEG2000, MPEG4, and a neural network-based method--for all of these different image types and often on nearly all of the individual images.
****

Listen to Look: Action Recognition by Previewing Audio
Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, Lorenzo Torresani; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10457-10467
In the face of the video data deluge, today's expensive clip-level classifiers are increasingly impractical. We propose a framework for efficient action recognition in untrimmed video that uses audio as a preview mechanism to eliminate both short-term and long-term visual redundancies. First, we devise an ImgAud2Vid framework that hallucinates clip-level features by distilling from lighter modalities---a single frame and its accompanying audio---reducing short-term temporal redundancy for efficient clip-level recognition. Second, building on ImgAud2Vid, we further propose ImgAud-Skimming, an attention-based long short-term memory network that iteratively selects useful moments in untrimmed videos, reducing long-term temporal redundancy for efficient video-level recognition. Extensive experiments on four action recognition datasets demonstrate that our method achieves the state-of-the-art in terms of both recognition accuracy and speed.
****

Memory Enhanced Global-Local Aggregation for Video Object Detection
Yihong Chen, Yue Cao, Han Hu, Liwei Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10337-10346
How do humans recognize an object in a piece of video? Due to the deteriorated quality of single frame, it may be hard for people to identify an occluded object in this frame by just utilizing information within one image. We argue that there are two important cues for humans to recognize objects in videos: the global semantic information and the local localization information. Recently, plenty of methods adopt the self-attention mechanisms to enhance the features in key frame with either global semantic information or local localization information. In this paper we introduce memory enhanced global-local aggregation (MEGA) network, which is among the first trials that takes full consideration of both global and local information. Furthermore, empowered by a novel and carefully-designed Long Range Memory (LRM) module, our proposed MEGA could enable the key frame to get access to much more content than any previous methods. Enhanced by these two sources of information, our method achieves state-of-the-art performance on ImageNet VID dataset. Code is available at https://github.com/Scalsol/mega.pytorch.
****

Self-Training With Noisy Student Improves ImageNet Classification
Qizhe Xie, Minh-Thang Luong, Eduard Hovy, Quoc V. Le; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10687-10698
We present a simple self-training method that achieves 88.4% top-1 accuracy on ImageNet, which is 2.0% better than the state-of-the-art model that requires 3.5B weakly labeled Instagram images. On robustness test sets, it improves ImageNet-A top-1 accuracy from 61.0% to 83.7%, reduces ImageNet-C mean corruption error from 45.7 to 28.3, and reduces ImageNet-P mean flip rate from 27.8 to 12.2. To achieve this result, we first train an EfficientNet model on labeled ImageNet images and use it as a teacher to generate pseudo labels on 300M unlabeled images. We then train a larger EfficientNet as a student model on the combination of labeled and pseudo labeled images. We iterate this process by putting back the student as the teacher. During the generation of the pseudo labels, the teacher is not noised so that the pseudo labels are as accurate as possible. However, during the learning of the student, we inject noise such as dropout, stochastic depth and data augmentation via RandAugment to the student so that the student generali

zes better than the teacher.
****
Distilling Cross-Task Knowledge via Relationship Matching
Han-Jia Ye, Su Lu, De-Chuan Zhan; Proceedings of the IEEE/CVF Conference on Co mputer Vision and Pattern Recognition (CVPR), 2020, pp. 12396-12405
The discriminative knowledge from a high-capacity deep neural network (a.k.a. th e "teacher") could be distilled to facilitate the learning efficacy of a shallow counterpart (a.k.a. the "student"). This paper deals with a general scenario re using the knowledge from a cross-task teacher --- two models are targeting non-o verlapping label spaces. We emphasize that the comparison ability between instan ces acts as an essential factor threading knowledge across domains, and propose the RElationship FacIlitated Local cLassifiEr Distillation (ReFilled) approach, which decomposes the knowledge distillation flow into branches for embedding and the top-layer classifier. In particular, different from reconciling the instanc e-label confidence between models, ReFilled requires the teacher to reweight the hard triplets push forwarded by the student so that the similarity comparison l evels between instances are matched. A local embedding-induced classifier from t he teacher further supervises the student's classification confidence. ReFilled demonstrates its effectiveness when reusing cross-task models, and also achieves state-of-the-art performance on the standard knowledge distillation benchmarks. The code of the paper can be accessed at https://github.com/njulus/ReFilled.
****
Multi-Modal Graph Neural Network for Joint Reasoning on Vision and Scene Text
Difei Gao, Ke Li, Ruiping Wang, Shiguang Shan, Xilin Chen; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, p p. 12746-12756
Answering questions that require reading texts in an image is challenging for cu rrent models. One key difficulty of this task is that rare, polysemous, and ambi guous words frequently appear in images, e.g., names of places, products, and sp orts teams. To overcome this difficulty, only resorting to pre-trained word embe dding models is far from enough. A desired model should utilize the rich informa tion in multiple modalities of the image to help understand the meaning of scene texts, e.g., the prominent text on a bottle is most likely to be the brand. Fol lowing this idea, we propose a novel VQA approach, Multi-Modal Graph Neural Netw ork (MM-GNN). It first represents an image as a graph consisting of three sub-gr aphs, depicting visual, semantic, and numeric modalities respectively. Then, we introduce three aggregators which guide the message passing from one graph to an other to utilize the contexts in various modalities, so as to refine the feature s of nodes. The updated nodes have better features for the downstream question a nswering module. Experimental evaluations show that our MM-GNN represents the sc ene texts better and obviously facilitates the performances on two VQA tasks tha t require reading scene texts.
****
Detecting Adversarial Samples Using Influence Functions and Nearest Neighbors
Gilad Cohen, Guillermo Sapiro, Raja Giryes; Proceedings of the IEEE/CVF Confer ence on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14453-14462
Deep neural networks (DNNs) are notorious for their vulnerability to adversarial attacks, which are small perturbations added to their input images to mislead t heir prediction. Detection of adversarial examples is, therefore, a fundamental requirement for robust classification frameworks. In this work, we present a met hod for detecting such adversarial attacks, which is suitable for any pre-traine d neural network classifier. We use influence functions to measure the impact of every training sample on the validation set data. From the influence scores, we find the most supportive training samples for any given validation example. A k -nearest neighbor (k-NN) model fitted on the DNN's activation layers is employed to search for the ranking of these supporting training samples. We observe that these samples are highly correlated with the nearest neighbors of the normal in puts, while this correlation is much weaker for adversarial inputs. We train an adversarial detector using the k-NN ranks and distances and show that it success fully distinguishes adversarial examples, getting state-of-the-art results on si

x attack methods with three datasets. Code is available at https://github.com/giladcohen/NNIF_adv_defense.
****

LiDAR-Based Online 3D Video Object Detection With Graph-Based Message Passing and Spatiotemporal Transformer Attention
Junbo Yin, Jianbing Shen, Chenye Guan, Dingfu Zhou, Ruigang Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11495-11504
Existing LiDAR-based 3D object detectors usually focus on the single-frame detection, while ignoring the spatiotemporal information in consecutive point cloud frames. In this paper, we propose an end-to-end online 3D video object detector that operates on point cloud sequences. The proposed model comprises a spatial feature encoding component and a spatiotemporal feature aggregation component. In the former component, a novel Pillar Message Passing Network (PMPNet) is proposed to encode each discrete point cloud frame. It adaptively collects information for a pillar node from its neighbors by iterative message passing, which effectively enlarges the receptive field of the pillar feature. In the latter component, we propose an Attentive Spatiotemporal Transformer GRU (AST-GRU) to aggregate the spatiotemporal information, which enhances the conventional ConvGRU with an attentive memory gating mechanism. AST-GRU contains a Spatial Transformer Attention (STA) module and a Temporal Transformer Attention (TTA) module, which can emphasize the foreground objects and align the dynamic objects, respectively. Experimental results demonstrate that the proposed 3D video object detector achieves state-of-the-art performance on the large-scale nuScenes benchmark.
****

Iterative Context-Aware Graph Inference for Visual Dialog
Dan Guo, Hui Wang, Hanwang Zhang, Zheng-Jun Zha, Meng Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10055-10064
Visual dialog is a challenging task that requires the comprehension of the semantic dependencies among implicit visual and textual contexts. This task can refer to the relation inference in a graphical model with sparse contexts and unknown graph structure (relation descriptor), and how to model the underlying context-aware relation inference is critical. To this end, we propose a novel Context-Aware Graph (CAG) neural network. Each node in the graph corresponds to a joint semantic feature, including both object-based (visual) and history-related (textual) context representations. The graph structure (relations in dialog) is iteratively updated using an adaptive top-K message passing mechanism. Specifically, in every message passing step, each node selects the most K relevant nodes, and only receives messages from them. Then, after the update, we impose graph attention on all the nodes to get the final graph embedding and infer the answer. In CAG, each node has dynamic relations in the graph (different related K neighbor nodes), and only the most relevant nodes are attributive to the context-aware relational graph inference. Experimental results on VisDial v0.9 and v1.0 datasets show that CAG outperforms comparative methods. Visualization results further validate the interpretability of our method.
****

Unsupervised Person Re-Identification via Multi-Label Classification
Dongkai Wang, Shiliang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10981-10990
The challenge of unsupervised person re-identification (ReID) lies in learning discriminative features without true labels. This paper formulates unsupervised person ReID as a multi-label classification task to progressively seek true labels. Our method starts by assigning each person image with a single-class label, then evolves to multi-label classification by leveraging the updated ReID model for label prediction. The label prediction comprises similarity computation and cycle consistency to ensure the quality of predicted labels. To boost the ReID model training efficiency in multi-label classification, we further propose the memory-based multi-label classification loss (MMCL). MMCL works with memory-based non-parametric classifier and integrates multi-label classification and single-l

abel classification in an unified framework. Our label prediction and MMCL work iteratively and substantially boost the ReID performance. Experiments on several large-scale person ReID datasets demonstrate the superiority of our method in unsupervised person ReID. Our method also allows to use labeled person images in other domains. Under this transfer learning setting, our method also achieves state-of-the-art performance.
****

Hit-Detector: Hierarchical Trinity Architecture Search for Object Detection
Jianyuan Guo, Kai Han, Yunhe Wang, Chao Zhang, Zhaohui Yang, Han Wu, Xinghao Chen, Chang Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11405-11414
Neural Architecture Search (NAS) has achieved great success in image classification task. Some recent works have managed to explore the automatic design of efficient backbone or feature fusion layer for object detection. However, these methods focus on searching only one certain component of object detector while leaving others manually designed. We identify the inconsistency between searched component and manually designed ones would withhold the detector of stronger performance. To this end, we propose a hierarchical trinity search framework to simultaneously discover efficient architectures for all components (i.e. backbone, neck, and head) of object detector in an end-to-end manner. In addition, we empirically reveal that different parts of the detector prefer different operators. Motivated by this, we employ a novel scheme to automatically screen different sub search spaces for different components so as to perform the end-to-end search for each component on the corresponding sub search space efficiently. Without bells and whistles, our searched architecture, namely Hit-Detector, achieves 41.4% mAP on COCO minival set with 27M parameters. Our implementation is available at \href https://github.com/ggjy/HitDet.pytorch https://github.com/ggjy/HitDet.pytorch .
****

Visual-Semantic Matching by Exploring High-Order Attention and Distraction
Yongzhi Li, Duo Zhang, Yadong Mu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12786-12795
Cross-modality semantic matching is a vital task in computer vision and has attracted increasing attention in recent years. Existing methods mainly explore object-based alignment between image objects and text words. In this work, we address this task from two previously-ignored aspects: high-order semantic information (e.g., object-predicate-subject triplet, object-attribute pair) and visual distraction (i.e., despite the high relevance to textual query, images may also contain many prominent distracting objects or visual relations). Specifically, we build scene graphs for both visual and textual modalities. Our technical contributions are two-folds: firstly, we formulate the visual-semantic matching task as an attention-driven cross-modality scene graph matching problem. Graph convolutional networks (GCNs) are used to extract high-order information from two scene graphs. A novel cross-graph attention mechanism is proposed to contextually reweigh graph elements and calculate the inter-graph similarity; Secondly, some top-ranked samples are indeed false matching due to the co-occurrence of both highly-relevant and distracting information. We devise an information-theoretic measure for estimating semantic distraction and re-ranking the initial retrieval results. Comprehensive experiments and ablation studies on two large public datasets (MS-COCO and Flickr30K) demonstrate the superiority of the proposed method and the effectiveness of both high-order attention and distraction.
****

Disparity-Aware Domain Adaptation in Stereo Image Restoration
Bo Yan, Chenxi Ma, Bahetiyaer Bare, Weimin Tan, Steven C. H. Hoi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13179-13187
Under stereo settings, the problems of disparity estimation, stereo magnification and stereo-view synthesis have gathered wide attention. However, the limited image quality brings non-negligible difficulties in developing related applications and becomes the main bottleneck of stereo images. To the best of our knowledg

e, stereo image restoration is rarely studied. Towards this end, this paper analyses how to effectively explore disparity information, and proposes a unified stereo image restoration framework. The proposed framework explicitly learn the inherent pixel correspondence between stereo views and restores stereo image with the cross-view information at image and feature level. A Feature Modulation Dense Block (FMDB) is introduced to insert disparity prior throughout the whole network. The experiments in terms of efficiency, objective and perceptual quality, and the accuracy of depth estimation demonstrates the superiority of the proposed framework on various stereo image restoration tasks.
****

Assessing Eye Aesthetics for Automatic Multi-Reference Eye In-Painting
Bo Yan, Qing Lin, Weimin Tan, Shili Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13509-13517
With the wide use of artistic images, aesthetic quality assessment has been widely concerned. How to integrate aesthetics into image editing is still a problem worthy of discussion. In this paper, aesthetic assessment is introduced into eye in-painting task for the first time. We construct an eye aesthetic dataset, and train the eye aesthetic assessment network on this basis. Then we propose a novel eye aesthetic and face semantic guided multi-reference eye inpainting GAN approach (AesGAN), which automatically selects the best reference under the guidance of eye aesthetics. A new aesthetic loss has also been introduced into the network to learn the eye aesthetic features and generate highquality eyes. We prove the effectiveness of eye aesthetic assessment in our experiments, which may inspire more applications of aesthetics assessment. Both qualitative and quantitative experimental results show that the proposed AesGAN can produce more natural and visually attractive eyes compared with state-of-the-art methods.
****

Equalization Loss for Long-Tailed Object Recognition
Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, Junjie Yan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11662-11671
Object recognition techniques using convolutional neural networks (CNN) have achieved great success. However, state-of-the-art object detection methods still perform poorly on large vocabulary and long-tailed datasets, e.g. LVIS. In this work, we analyze this problem from a novel perspective: each positive sample of one category can be seen as a negative sample for other categories, making the tail categories receive more discouraging gradients. Based on it, we propose a simple but effective loss, named equalization loss, to tackle the problem of long-tailed rare categories by simply ignoring those gradients for rare categories. The equalization loss protects the learning of rare categories from being at a disadvantage during the network parameter updating. Thus the model is capable of learning better discriminative features for objects of rare classes. Without any bells and whistles, our method achieves AP gains of 4.1% and 4.8% for the rare and common categories on the challenging LVIS benchmark, compared to the Mask R-CNN baseline. With the utilization of the effective equalization loss, we finally won the 1st place in the LVIS Challenge 2019. Code has been made available at: https://github.com/tztztztztz/eql.detectron2
****

Sideways: Depth-Parallel Training of Video Models
Mateusz Malinowski, Grzegorz Swirszcz, Joao Carreira, Viorica Patraucean; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11834-11843
We propose Sideways, an approximate backpropagation scheme for training video models. In standard backpropagation, the gradients and activations at every computation step through the model are temporally synchronized. The forward activations need to be stored until the backward pass is executed, preventing inter-layer (depth) parallelization. However, can we leverage smooth, redundant input streams such as videos to develop a more efficient training scheme? Here, we explore an alternative to backpropagation; we overwrite network activations whenever new ones, i.e., from new frames, become available. Such a more gradual accumulation

of information from both passes breaks the precise correspondence between gradients and activations, leading to theoretically more noisy weight updates. Counter-intuitively, we show that Sideways training of deep convolutional video networks not only still converges, but can also potentially exhibit better generalization compared to standard synchronized backpropagation.
****

Hierarchical Conditional Relation Networks for Video Question Answering
Thao Minh Le, Vuong Le, Svetha Venkatesh, Truyen Tran; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9972-9981
Video question answering (VideoQA) is challenging as it requires modeling capacity to distill dynamic visual artifacts and distant relations and to associate them with linguistic concepts. We introduce a general-purpose reusable neural unit called Conditional Relation Network (CRN) that serves as a building block to construct more sophisticated structures for representation and reasoning over video. CRN takes as input an array of tensorial objects and a conditioning feature, and computes an array of encoded output objects. Model building becomes a simple exercise of replication, rearrangement and stacking of these reusable units for diverse modalities and contextual information. This design thus supports high-order relational and multi-step reasoning. The resulting architecture for VideoQA is a CRN hierarchy whose branches represent sub-videos or clips, all sharing the same question as the contextual condition. Our evaluations on well-known datasets achieved new SoTA results, demonstrating the impact of building a general-purpose reasoning unit on complex domains such as VideoQA.
****

RankMI: A Mutual Information Maximizing Ranking Loss
Mete Kemertas, Leila Pishdad, Konstantinos G. Derpanis, Afsaneh Fazly; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14362-14371
We introduce an information-theoretic loss function, RankMI, and an associated training algorithm for deep representation learning for image retrieval. Our proposed framework consists of alternating updates to a network that estimates the divergence between distance distributions of matching and non-matching pairs of learned embeddings, and an embedding network that maximizes this estimate via sampled negatives. In addition, under this information-theoretic lens we draw connections between RankMI and commonly-used ranking losses, e.g., triplet loss. We extensively evaluate RankMI on several standard image retrieval datasets, namely, CUB-200-2011, CARS-196, and Stanford Online Products. Our method achieves competitive results or significant improvements over previous reported results on all datasets.
****

HAMBox: Delving Into Mining High-Quality Anchors on Face Detection
Yang Liu, Xu Tang, Junyu Han, Jingtuo Liu, Dinger Rui, Xiang Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13046-13054
Current face detectors utilize anchors to frame a multi-task learning problem which combines classification and bounding box regression. Effective anchor design and anchor matching strategy enable face detectors to localize faces under large pose and scale variations. However, we observe that, more than 80% correctly predicted bounding boxes are regressed from the unmatched anchors (the IoUs between anchors and target faces are lower than a threshold) in the inference phase. It indicates that these unmatched anchors perform excellent regression ability, but the existing methods neglect to learn from them. In this paper, we propose an Online High-quality Anchor Mining Strategy (HAMBox), which explicitly helps outer faces compensate with high-quality anchors. Our proposed HAMBox method could be a general strategy for anchor-based single-stage face detection. Experiments on various datasets, including WIDER FACE, FDDB, AFW and PASCAL Face, demonstrate the superiority of the proposed method.
****

McFlow: Monte Carlo Flow Models for Data Imputation

Trevor W. Richardson, Wencheng Wu, Lei Lin, Beilei Xu, Edgar A. Bernal; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14205-14214

We consider the topic of data imputation, a foundational task in machine learning that addresses issues with missing data. To that end, we propose MCFlow, a deep framework for imputation that leverages normalizing flow generative models and Monte Carlo sampling. We address the causality dilemma that arises when training models with incomplete data by introducing an iterative learning scheme which alternately updates the density estimate and the values of the missing entries in the training data. We provide extensive empirical validation of the effectiveness of the proposed method on standard multivariate and image datasets, and benchmark its performance against state-of-the-art alternatives. We demonstrate that MCFlow is superior to competing methods in terms of the quality of the imputed data, as well as with regards to its ability to preserve the semantic structure of the data.

****

MonoPair: Monocular 3D Object Detection Using Pairwise Spatial Relationships

Yongjian Chen, Lei Tai, Kai Sun, Mingyang Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12093-12102

Monocular 3D object detection is an essential component in autonomous driving while challenging to solve, especially for those occluded samples which are only partially visible. Most detectors consider each 3D object as an independent training target, inevitably resulting in a lack of useful information for occluded samples. To this end, we propose a novel method to improve the monocular 3D object detection by considering the relationship of paired samples. This allows us to encode spatial constraints for partially-occluded objects from their adjacent neighbors. Specifically, the proposed detector computes uncertainty-aware predictions for object locations and 3D distances for the adjacent object pairs, which are subsequently jointly optimized by nonlinear least squares. Finally, the one-stage uncertainty-aware prediction structure and the post-optimization module are dedicatedly integrated for ensuring the run-time efficiency. Experiments demonstrate that our method yields the best performance on KITTI 3D detection benchmark, by outperforming state-of-the-art competitors by wide margins, especially for the hard samples.

****

KeyPose: Multi-View 3D Labeling and Keypoint Estimation for Transparent Objects

Xingyu Liu, Rico Jonschkowski, Anelia Angelova, Kurt Konolige; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11602-11610

Estimating the 3D pose of desktop objects is crucial for applications such as robotic manipulation. Many existing approaches to this problem require a depth map of the object for both training and prediction, which restricts them to opaque, lambertian objects that produce good returns in an RGBD sensor. In this paper we forgo using a depth sensor in favor of raw stereo input. We address two problems: first, we establish an easy method for capturing and labeling 3D keypoints on desktop objects with an RGB camera; and second, we develop a deep neural network, called KeyPose, that learns to accurately predict object poses using 3D keypoints, from stereo input, and works even for transparent objects. To evaluate the performance of our method, we create a dataset of 15 clear objects in five classes, with 48K 3D-keypoint labeled images. We train both instance and category models, and show generalization to new textures, poses, and objects. KeyPose surpasses state-of-the-art performance in 3D pose estimation on this dataset by factors of 1.5 to 3.5, even in cases where the competing method is provided with ground-truth depth. Stereo input is essential for this performance as it improves results compared to using monocular input by a factor of 2. We will release a public version of the data capture and labeling pipeline, the transparent object database, and the KeyPose models and evaluation code. Project website: https://sites.google.com/corp/view/keypose.

****

Putting Visual Object Recognition in Context

Mengmi Zhang, Claire Tseng, Gabriel Kreiman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12985-12994
Context plays an important role in visual recognition. Recent studies have shown that visual recognition networks can be fooled by placing objects in inconsistent contexts (e.g., a cow in the ocean). To model the role of contextual information in visual recognition, we systematically investigated ten critical properties of where, when, and how context modulates recognition, including the amount of context, context and object resolution, geometrical structure of context, context congruence, and temporal dynamics of contextual modulation. The tasks involved recognizing a target object surrounded with context in a natural image. As an essential benchmark, we conducted a series of psychophysics experiments where we altered one aspect of context at a time, and quantified recognition accuracy. We propose a biologically-inspired context-aware object recognition model consisting of a two-stream architecture. The model processes visual information at the fovea and periphery in parallel, dynamically incorporates object and contextual information, and sequentially reasons about the class label for the target object. Across a wide range of behavioral tasks, the model approximates human level performance without retraining for each task, captures the dependence of context enhancement on image properties, and provides initial steps towards integrating scene and object information for visual recognition. All source code and data are publicly available: https://github.com/kreimanlab/Put-In-Context.
****

Multi-Path Learning for Object Pose Estimation Across Domains
Martin Sundermeyer, Maximilian Durner, En Yen Puang, Zoltan-Csaba Marton, Narunas Vaskevicius, Kai O. Arras, Rudolph Triebel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13916-13925
We introduce a scalable approach for object pose estimation trained on simulated RGB views of multiple 3D models together. We learn an encoding of object views that does not only describe an implicit orientation of all objects seen during training, but can also relate views of untrained objects. Our single-encoder-multi-decoder network is trained using a technique we denote "multi-path learning": While the encoder is shared by all objects, each decoder only reconstructs views of a single object. Consequently, views of different instances do not have to be separated in the latent space and can share common features. The resulting encoder generalizes well from synthetic to real data and across various instances, categories, model types and datasets. We systematically investigate the learned encodings, their generalization, and iterative refinement strategies on the ModelNet40 and T-LESS dataset. Despite training jointly on multiple objects, our 6D Object Detection pipeline achieves state-of-the-art results on T-LESS at much lower runtimes than competing approaches.
****

Instance Credibility Inference for Few-Shot Learning
Yikai Wang, Chengming Xu, Chen Liu, Li Zhang, Yanwei Fu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12836-12845
Few-shot learning (FSL) aims to recognize new objects with extremely limited training data for each category. Previous efforts are made by either leveraging meta-learning paradigm or novel principles in data augmentation to alleviate this extremely data-scarce problem. In contrast, this paper presents a simple statistical approach, dubbed Instance Credibility Inference (ICI) to exploit the distribution support of unlabeled instances for few-shot learning. Specifically, we first train a linear classifier with the labeled few-shot examples and use it to infer the pseudo-labels for the unlabeled data. To measure the credibility of each pseudo-labeled instance, we then propose to solve another linear regression hypothesis by increasing the sparsity of the incidental parameters and rank the pseudo-labeled instances with their sparsity degree. We select the most trustworthy pseudo-labeled instances alongside the labeled examples to re-train the linear classifier. This process is iterated until all the unlabeled samples are included in the expanded training set, i.e. the pseudo-label is converged for unlabeled

data pool. Extensive experiments under two few-shot settings show that our simple approach can establish new state-of-the-arts on four widely used few-shot learning benchmark datasets including miniImageNet, tieredImageNet, CIFAR-FS, and CUB. Our code is available at: https://github.com/Yikai-Wang/ICI-FSL
****

From Paris to Berlin: Discovering Fashion Style Influences Around the World
Ziad Al-Halah, Kristen Grauman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10136-10145
The evolution of clothing styles and their migration across the world is intriguing, yet difficult to describe quantitatively. We propose to discover and quantify fashion influences from everyday images of people wearing clothes. We introduce an approach that detects which cities influence which other cities in terms of propagating their styles. We then leverage the discovered influence patterns to inform a forecasting model that predicts the popularity of any given style at any given city into the future. Demonstrating our idea with GeoStyle--a large-scale dataset of 7.7M images covering 44 major world cities, we present the discovered influence relationships, revealing how cities exert and receive fashion influence for an array of 50 observed visual styles. Furthermore, the proposed forecasting model achieves state-of-the-art results for a challenging style forecasting task, showing the advantage of grounding visual style evolution both spatially and temporally.
****

Severity-Aware Semantic Segmentation With Reinforced Wasserstein Training
Xiaofeng Liu, Wenxuan Ji, Jane You, Georges El Fakhri, Jonghye Woo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12566-12575
Semantic segmentation is a class of methods to classify each pixel in an image into semantic classes, which is critical for autonomous vehicles and surgery systems. Cross-entropy (CE) loss-based deep neural networks (DNN) achieved great success w.r.t. the accuracy-based metrics, e.g., mean Intersection-over Union. However, the CE loss has a limitation in that it ignores varying degrees of severity of pair-wise misclassified results. For instance, classifying a car into the road is much more terrible than recognizing it as a bus. To sidestep this, in this work, we propose to incorporate the severity-aware inter-class correlation into our Wasserstein training framework by configuring its ground distance matrix. In addition, our method can adaptively learn the ground metric in a high-fidelity simulator, following a reinforcement alternative optimization scheme. We evaluate our method using the CARLA simulator with the Deeplab backbone, demonstraing that our method significantly improves the survival time in the CARLA simulator. In addition, our method can be readily applied to existing DNN architectures and algorithms while yielding superior performance. We report results from experiments carried out with the CamVid and Cityscapes datasets.
****

Sketchformer: Transformer-Based Representation for Sketched Structure
Leo Sampaio Ferraz Ribeiro, Tu Bui, John Collomosse, Moacir Ponti; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14153-14162
Sketchformer is a novel transformer-based representation for encoding free-hand sketches input in a vector form, i.e. as a sequence of strokes. Sketchformer effectively addresses multiple tasks: sketch classification, sketch based image retrieval (SBIR), and the reconstruction and interpolation of sketches. We report several variants exploring continuous and tokenized input representations, and contrast their performance. Our learned embedding, driven by a dictionary learning tokenization scheme, yields state of the art performance in classification and image retrieval tasks, when compared against baseline representations driven by LSTM sequence to sequence architectures: SketchRNN and derivatives. We show that sketch reconstruction and interpolation are improved significantly by the Sketchformer embedding for complex sketches with longer stroke sequences.
****

Detail-recovery Image Deraining via Context Aggregation Networks

Sen Deng, Mingqiang Wei, Jun Wang, Yidan Feng, Luming Liang, Haoran Xie, F u Lee Wang, Meng Wang; Proceedings of the IEEE/CVF Conference on Computer Visio n and Pattern Recognition (CVPR), 2020, pp. 14560-14569
This paper looks at this intriguing question: are single images with their detai ls lost during deraining, reversible to their artifact-free status? We propose a n end-to-end detail-recovery image deraining network (termed a DRDNet) to solve the problem. Unlike existing image deraining approaches that attempt to meet the conflicting goal of simultaneously deraining and preserving details in a unifie d framework, we propose to view rain removal and detail recovery as two seperate tasks, so that each part could specialize rather than trade-off between two con flicting goals. Specifically, we introduce two parallel sub-networks with a comp rehensive loss function which synergize to derain and recover the lost details c aused by deraining. For complete rain removal, we present a rain residual networ k with the squeeze-and-excitation (SE) operation to remove rain streaks from the rainy images. For detail recovery, we construct a specialized detail repair net work consisting of welldesigned blocks, named structure detail context aggregati on block (SDCAB), to encourage the lost details to return for eliminating image degradations. Moreover, the detail recovery branch of our proposed detail repair framework is detachable and can be incorporated into existing deraining methods to boost their performances. DRD-Net has been validated on several well-known b enchmark datasets in terms of deraining robustness and detail accuracy. Comparis ons show clear visual and numerical improvements of our method over the state-of -the-arts.
****

Dynamic Refinement Network for Oriented and Densely Packed Object Detection
Xingjia Pan, Yuqiang Ren, Kekai Sheng, Weiming Dong, Haolei Yuan, Xiaowei G uo, Chongyang Ma, Changsheng Xu; Proceedings of the IEEE/CVF Conference on Com puter Vision and Pattern Recognition (CVPR), 2020, pp. 11207-11216
Object detection has achieved remarkable progress in the past decade. However, t he detection of oriented and densely packed objects remains challenging because of following inherent reasons: (1) receptive fields of neurons are all axis-alig ned and of the same shape, whereas objects are usually of diverse shapes and ali gn along various directions; (2) detection models are typically trained with gen eric knowledge and may not generalize well to handle specific objects at test ti me; (3) the limited dataset hinders the development on this task. To resolve the first two issues, we present a dynamic refinement network that consists of two novel components, i.e., a feature selection module (FSM) and a dynamic refinemen t head (DRH). Our FSM enables neurons to adjust receptive fields in accordance w ith the shapes and orientations of target objects, whereas the DRH empowers our model to refine the prediction dynamically in an object-aware manner. To address the limited availability of related benchmarks, we collect an extensive and ful ly annotated dataset, namely, SKU110K-R, which is relabeled with oriented boundi ng boxes based on SKU110K. We perform quantitative evaluations on several public ly available benchmarks including DOTA, HRSC2016, SKU110K, and our own SKU110K-R dataset. Experimental results show that our method achieves consistent and subs tantial gains compared with baseline approaches. Our source code and dataset wil l be released to encourage follow-up research.
****

Self-Trained Deep Ordinal Regression for End-to-End Video Anomaly Detection
Guansong Pang, Cheng Yan, Chunhua Shen, Anton van den Hengel, Xiao Bai; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2020, pp. 12173-12182
Video anomaly detection is of critical practical importance to a variety of real applications because it allows human attention to be focused on events that are likely to be of interest, in spite of an otherwise overwhelming volume of video . We show that applying self-trained deep ordinal regression to video anomaly de tection overcomes two key limitations of existing methods, namely, 1) being high ly dependent on manually labeled normal training data; and 2) sub-optimal featur e learning. By formulating a surrogate two-class ordinal regression task we devi se an end-to-end trainable video anomaly detection approach that enables joint r

epresentation learning and anomaly scoring without manually labeled normal/abnormal data. Experiments on eight real-world video scenes show that our proposed method outperforms state-of-the-art methods that require no labeled training data by a substantial margin, and enables easy and accurate localization of the identified anomalies. Furthermore, we demonstrate that our method offers effective human-in-the-loop anomaly detection which can be critical in applications where anomalies are rare and the false-negative cost is high.
****

Smoothing Adversarial Domain Attack and P-Memory Reconsolidation for Cross-Domain Person Re-Identification
Guangcong Wang, Jian-Huang Lai, Wenqi Liang, Guangrun Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10568-10577
Most of the existing person re-identification (re-ID) methods achieve promising accuracy in a supervised manner, but they assume the identity labels of the target domain is available. This greatly limits the scalability of person re-ID in real-world scenarios. Therefore, the current person re-ID community focuses on the cross-domain person re-ID that aims to transfer the knowledge from a labeled source domain to an unlabeled target domain and exploits the specific knowledge from the data distribution of the target domain to further improve the performance. To reduce the gap between the source and target domains, we propose a Smoothing Adversarial Domain Attack (SADA) approach that guides the source domain images to align the target domain images by using a trained camera classifier. To stabilize a memory trace of cross-domain knowledge transfer after its initial acquisition from the source domain, we propose a p-Memory Reconsolidation (pMR) method that reconsolidates the source knowledge with a small probability p during the self-training of the target domain. With both SADA and pMR, the proposed method significantly improves the cross-domain person re-ID. Extensive experiments on Market-1501 and DukeMTMC-reID benchmarks show that our pMR-SADA outperforms all of the state-of-the-arts by a large margin.
****

Predicting Sharp and Accurate Occlusion Boundaries in Monocular Depth Estimation Using Displacement Fields
Michael Ramamonjisoa, Yuming Du, Vincent Lepetit; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14648-14657
Current methods for depth map prediction from monocular images tend to predict smooth, poorly localized contours for the occlusion boundaries in the input image. This is unfortunate as occlusion boundaries are important cues to recognize objects, and as we show, may lead to a way to discover new objects from scene reconstruction. To improve predicted depth maps, recent methods rely on various forms of filtering or predict an additive residual depth map to refine a first estimate. We instead learn to predict, given a depth map predicted by some reconstruction method, a 2D displacement field able to re-sample pixels around the occlusion boundaries into sharper reconstructions. Our method can be applied to the output of any depth estimation method, in an end-to-end trainable fashion. For evaluation, we manually annotated the occlusion boundaries in all the images in the test split of popular NYUv2-Depth dataset. We show that our approach improves the localization of occlusion boundaries for all state-of-the-art monocular depth estimation methods that we could evaluate, without degrading the depth accuracy for the rest of the images.
****

Spatiotemporal Fusion in 3D CNNs: A Probabilistic View
Yizhou Zhou, Xiaoyan Sun, Chong Luo, Zheng-Jun Zha, Wenjun Zeng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9829-9838
Despite the success in still image recognition, deep neural networks for spatiotemporal signal tasks (such as human action recognition in videos) still suffers from low efficacy and inefficiency over the past years. Recently, human experts have put more efforts into analyzing the importance of different components in 3

D convolutional neural networks (3D CNNs) to design more powerful spatiotemporal learning backbones. Among many others, spatiotemporal fusion is one of the essentials. It controls how spatial and temporal signals are extracted at each layer during inference. Previous attempts usually start by ad-hoc designs that empirically combine certain convolutions and then draw conclusions based on the performance obtained by training the corresponding networks. These methods only support network-level analysis on limited number of fusion strategies. In this paper, we propose to convert the spatiotemporal fusion strategies into a probability space, which allows us to perform network-level evaluations of various fusion strategies without having to train them separately. Besides, we can also obtain fine-grained numerical information such as layer-level preference on spatiotemporal fusion within the probability space. Our approach greatly boosts the efficiency of analyzing spatiotemporal fusion. Based on the probability space, we further generate new fusion strategies which achieve the state-of-the-art performance on four well-known action recognition datasets.

****

LiDARsim: Realistic LiDAR Simulation by Leveraging the Real World

Sivabalan Manivasagam, Shenlong Wang, Kelvin Wong, Wenyuan Zeng, Mikita Sazanovich, Shuhan Tan, Bin Yang, Wei-Chiu Ma, Raquel Urtasun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11167-11176

We tackle the problem of producing realistic simulations of LiDAR point clouds, the sensor of preference for most self-driving vehicles. We argue that, by leveraging real data, we can simulate the complex world more realistically compared to employing virtual worlds built from CAD/procedural models. Towards this goal, we first build a large catalog of 3D static maps and 3D dynamic objects by driving around several cities with our self-driving fleet. We can then generate scenarios by selecting a scene from our catalog and "virtually" placing the self-driving vehicle (SDV) and a set of dynamic objects from the catalog in plausible locations in the scene. To produce realistic simulations, we develop a novel simulator that captures both the power of physics-based and learning-based simulation. We first utilize raycasting over the 3D scene and then use a deep neural network to produce deviations from the physics-based simulation, producing realistic LiDAR point clouds. We showcase LiDARsim's usefulness for perception algorithms-testing on long-tail events and end-to-end closed-loop evaluation on safety-critical scenarios.

****

Counting Out Time: Class Agnostic Video Repetition Counting in the Wild

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, Andrew Zisserman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10387-10396

We present an approach for estimating the period with which an action is repeated in a video. The crux of the approach lies in constraining the period prediction module to use temporal self-similarity as an intermediate representation bottleneck that allows generalization to unseen repetitions in videos in the wild. We train this model, called RepNet, with a synthetic dataset that is generated from a large unlabeled video collection by sampling short clips of varying lengths and repeating them with different periods and counts. This combination of synthetic data and a powerful yet constrained model, allows us to predict periods in a class-agnostic fashion. Our model substantially exceeds the state of the art performance on existing periodicity (PERTUBE) and repetition counting (QUVA) benchmarks. We also collect a new challenging dataset called Countix (90 times larger than existing datasets) which captures the challenges of repetition counting in real-world videos. Project webpage: https://sites.google.com/view/repnet .

****

Inducing Hierarchical Compositional Model by Sparsifying Generator Network

Xianglei Xing, Tianfu Wu, Song-Chun Zhu, Ying Nian Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14296-14305

This paper proposes to learn hierarchical compositional AND-OR model for interpr

etable image synthesis by sparsifying the generator network. The proposed method adopts the scene-objects-parts-subparts-primitives hierarchy in image representation. A scene has different types (i.e., OR) each of which consists of a number of objects (i.e., AND). This can be recursively formulated across the scene-objects-parts-subparts hierarchy and is terminated at the primitive level (e.g., wavelets-like basis). To realize this AND-OR hierarchy in image synthesis, we learn a generator network that consists of the following two components: (i) Each layer of the hierarchy is represented by an over-completed set of convolutional basis functions. Off-the-shelf convolutional neural architectures are exploited to implement the hierarchy. (ii) Sparsity-inducing constraints are introduced in end-to-end training, which induces a sparsely activated and sparsely connected AND-OR model from the initially densely connected generator network. A straightforward sparsity-inducing constraint is utilized, that is to only allow the top-k basis functions to be activated at each layer (where k is a hyper-parameter). The learned basis functions are also capable of image reconstruction to explain the input images. In experiments, the proposed method is tested on four benchmark datasets. The results show that meaningful and interpretable hierarchical representations are learned with better qualities of image synthesis and reconstruction obtained than baselines.

****

What Deep CNNs Benefit From Global Covariance Pooling: An Optimization Perspective

Qilong Wang, Li Zhang, Banggu Wu, Dongwei Ren, Peihua Li, Wangmeng Zuo, Qinghua Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10771-10780

Recent works have demonstrated that global covariance pooling (GCP) has the ability to improve performance of deep convolutional neural networks (CNNs) on visual classification task. Despite considerable advance, the reasons on effectiveness of GCP on deep CNNs have not been well studied. In this paper, we make an attempt to understand what deep CNNs benefit from GCP in a viewpoint of optimization. Specifically, we explore the effect of GCP on deep CNNs in terms of the Lipschitzness of optimization loss and the predictiveness of gradients, and show that GCP can make the optimization landscape more smooth and the gradients more predictive. Furthermore, we discuss the connection between GCP and second-order optimization for deep CNNs. More importantly, above findings can account for several merits of covariance pooling for training deep CNNs that have not been recognized previously or fully explored, including significant acceleration of network convergence (i.e., the networks trained with GCP can support rapid decay of learning rates, achieving favorable performance while significantly reducing number of training epochs), stronger robustness to distorted examples generated by image corruptions and perturbations, and good generalization ability to different vision tasks, e.g., object detection and instance segmentation. We conduct extensive experiments using various deep CNN architectures on diversified tasks, and the results provide strong support to our findings.

****

EmotiCon: Context-Aware Multimodal Emotion Recognition Using Frege's Principle

Trisha Mittal, Pooja Guhan, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, Dinesh Manocha; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14234-14243

We present EmotiCon, a learning-based algorithm for context-aware perceived human emotion recognition from videos and images. Motivated by Frege's Context Principle from psychology, our approach combines three interpretations of context for emotion recognition. Our first interpretation is based on using multiple modalities (e.g.faces and gaits) for emotion recognition. For the second interpretation, we gather semantic context from the input image and use a self-attention-based CNN to encode this information. Finally, we use depth maps to model the third interpretation related to socio-dynamic interactions and proximity among agents. We demonstrate the efficiency of our network through experiments on EMOTIC, a benchmark dataset. We report an Average Precision (AP) score of 35.48 across 26 classes, which is an improvement of 7-8 over prior methods. We also introduce a n

ew dataset, GroupWalk, which is a collection of videos captured in multiple real-world settings of people walking. We report an AP of 65.83 across 4 categories on GroupWalk, which is also an improvement over prior methods.
****

Universal Weighting Metric Learning for Cross-Modal Matching
Jiwei Wei, Xing Xu, Yang Yang, Yanli Ji, Zheng Wang, Heng Tao Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13005-13014
Cross-modal matching has been a highlighted research topic in both vision and language areas. Learning appropriate mining strategy to sample and weight informative pairs is crucial for the cross-modal matching performance. However, most existing metric learning methods are developed for unimodal matching, which is unsuitable for cross-modal matching on multimodal data with heterogeneous features. To address this problem, we propose a simple and interpretable universal weighting framework for cross-modal matching, which provides a tool to analyze the interpretability of various loss functions. Furthermore, we introduce a new polynomial loss under the universal weighting framework, which defines a weight function for the positive and negative informative pairs respectively. Experimental results on two image-text matching benchmarks and two video-text matching benchmarks validate the efficacy of the proposed method.
****

Learning a Dynamic Map of Visual Appearance
Tawfiq Salem, Scott Workman, Nathan Jacobs; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12435-12444
The appearance of the world varies dramatically not only from place to place but also from hour to hour and month to month. Every day billions of images capture this complex relationship, many of which are associated with precise time and location metadata. We propose to use these images to construct a global-scale, dynamic map of visual appearance attributes. Such a map enables fine-grained understanding of the expected appearance at any geographic location and time. Our approach integrates dense overhead imagery with location and time metadata into a general framework capable of mapping a wide variety of visual attributes. A key feature of our approach is that it requires no manual data annotation. We demonstrate how this approach can support various applications, including image-driven mapping, image geolocalization, and metadata verification.
****

Learning From Synthetic Animals
Jiteng Mu, Weichao Qiu, Gregory D. Hager, Alan L. Yuille; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12386-12395
Despite great success in human parsing, progress for parsing other deformable articulated objects, like animals, is still limited by the lack of labeled data. In this paper, we use synthetic images and ground truth generated from CAD animal models to address this challenge. To bridge the domain gap between real and synthetic images, we propose a novel consistency-constrained semi-supervised learning method (CC-SSL). Our method leverages both spatial and temporal consistencies, to bootstrap weak models trained on synthetic data with unlabeled real images. We demonstrate the effectiveness of our method on highly deformable animals, such as horses and tigers. Without using any real image label, our method allows for accurate keypoint prediction on real images. Moreover, we quantitatively show that models using synthetic data achieve better generalization performance than models trained on real images across different domains in the Visual Domain Adaptation Challenge dataset. Our synthetic dataset contains 10+ animals with diverse poses and rich ground truth, which enables us to use the multi-task learning strategy to further boost models' performance.
****

3D Part Guided Image Editing for Fine-Grained Object Understanding
Zongdai Liu, Feixiang Lu, Peng Wang, Hui Miao, Liangjun Zhang, Ruigang Yang, Bin Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11336-11345

Holistically understanding an object with its 3D movable parts is essential for visual models of a robot to interact with the world. For example, only by understanding many possible part dynamics of other vehicles (e.g., door or trunk opening, taillight blinking for changing lane), a self-driving vehicle can be successful in dealing with emergency cases. However, existing visual models tackle rarely on these situations, but focus on bounding box detection. In this paper, we fill this important missing piece in autonomous driving by solving two critical issues. First, for dealing with data scarcity, we propose an effective training data generation process by fitting a 3D car model with dynamic parts to cars in real images. This allows us to directly edit the real images using the aligned 3D parts, yielding effective training data for learning robust deep neural networks (DNNs). Secondly, to benchmark the quality of 3D part understanding, we collected a large dataset in real driving scenario with cars in uncommon states (CUS), i.e. with door or trunk opened etc., which demonstrates that our trained network with edited images largely outperforms other baselines in terms of 2D detection and instance segmentation accuracy.

****

Learning Unseen Concepts via Hierarchical Decomposition and Composition

Muli Yang, Cheng Deng, Junchi Yan, Xianglong Liu, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10248-10256

Composing and recognizing new concepts from known sub-concepts has been a fundamental and challenging vision task, mainly due to 1) the diversity of sub-concepts and 2) the intricate contextuality between sub-concepts and their corresponding visual features. However, most of the current methods simply treat the contextuality as rigid semantic relationships and fail to capture fine-grained contextual correlations. We propose to learn unseen concepts in a hierarchical decomposition-and-composition manner. Considering the diversity of sub-concepts, our method decomposes each seen image into visual elements according to its labels, and learns corresponding sub-concepts in their individual subspaces. To model intricate contextuality between sub-concepts and their visual features, compositions are generated from these subspaces in three hierarchical forms, and the composed concepts are learned in a unified composition space. To further refine the captured contextual relationships, adaptively semi-positive concepts are defined and then learned with pseudo supervision exploited from the generated compositions. We validate the proposed approach on two challenging benchmarks, and demonstrate its superiority over state-of-the-art approaches.

****

Multi-Modality Cross Attention Network for Image and Sentence Matching

Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, Feng Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10941-10950

The key of image and sentence matching is to accurately measure the visual-semantic similarity between an image and a sentence. However, most existing methods make use of only the intra-modality relationship within each modality or the inter-modality relationship between image regions and sentence words for the cross-modal matching task. Different from them, in this work, we propose a novel MultiModality Cross Attention (MMCA) Network for image and sentence matching by jointly modeling the intra-modality and inter-modality relationships of image regions and sentence words in a unified deep model. In the proposed MMCA, we design a novel cross-attention mechanism, which is able to exploit not only the intra-modality relationship within each modality, but also the inter-modality relationship between image regions and sentence words to complement and enhance each other for image and sentence matching. Extensive experimental results on two standard benchmarks including Flickr30K and MS-COCO demonstrate that the proposed model performs favorably against state-of-the-art image and sentence matching methods.

****

Self-Supervised Domain-Aware Generative Network for Generalized Zero-Shot Learning

Jiamin Wu, Tianzhu Zhang, Zheng-Jun Zha, Jiebo Luo, Yongdong Zhang, Feng Wu

; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit ion (CVPR), 2020, pp. 12767-12776
Generalized Zero-Shot Learning (GZSL) aims at recognizing both seen and unseen c lasses by constructing correspondence between visual and semantic embedding. How ever, existing methods have severely suffered from the strong bias problem, wher e unseen instances in target domain tend to be recognized as seen classes in sou rce domain. To address this issue, we propose an end-to-end Self-supervised Doma in-aware Generative Network (SDGN) by integrating self-supervised learning into feature generating model for unbiased GZSL. The proposed SDGN model enjoys sever al merits. First, we design a cross-domain feature generating module to synthesi ze samples with high fidelity based on class embeddings, which involves a novel target domain discriminator to preserve the domain consistency. Second, we propo se a self-supervised learning module to investigate inter-domain relationships, where a set of anchors are introduced as a bridge between seen and unseen catego ries. In the shared space, we pull the distribution of target domain away from s ource domain, and obtain domain-aware features with high discriminative power fo r both seen and unseen classes. To our best knowledge, this is the first work to introduce self-supervised learning into GZSL as a learning guidance. Extensive experimental results on five standard benchmarks demonstrate that our model perf orms favorably against state-of-the-art GZSL methods.
****

EPOS: Estimating 6D Pose of Objects With Symmetries
Tomas Hodan, Daniel Barath, Jiri Matas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11703-11712
We present a new method for estimating the 6D pose of rigid objects with availab le 3D models from a single RGB input image. The method is applicable to a broad range of objects, including challenging ones with global or partial symmetries. An object is represented by compact surface fragments which allow handling symme tries in a systematic manner. Correspondences between densely sampled pixels and the fragments are predicted using an encoder-decoder network. At each pixel, th e network predicts: (i) the probability of each object's presence, (ii) the prob ability of the fragments given the object's presence, and (iii) the precise 3D l ocation on each fragment. A data-dependent number of corresponding 3D locations is selected per pixel, and poses of possibly multiple object instances are estim ated using a robust and efficient variant of the PnP-RANSAC algorithm. In the BO P Challenge 2019, the method outperforms all RGB and most RGB-D and D methods on the T-LESS and LM-O datasets. On the YCB-V dataset, it is superior to all compe titors, with a large margin over the second-best RGB method. Source code is at: cmp.felk.cvut.cz/epos.
****

Object Relational Graph With Teacher-Recommended Learning for Video Captioning
Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, Zhe ng-Jun Zha; Proceedings of the IEEE/CVF Conference on Computer Vision and Patter n Recognition (CVPR), 2020, pp. 13278-13288
Taking full advantage of the information from both vision and language is critic al for the video captioning task. Existing models lack adequate visual represent ation due to the neglect of interaction between object, and sufficient training for content-related words due to long-tailed problems. In this paper, we propose a complete video captioning system including both a novel model and an effectiv e training strategy. Specifically, we propose an object relational graph (ORG) b ased encoder, which captures more detailed interaction features to enrich visual representation. Meanwhile, we design a teacher-recommended learning (TRL) metho d to make full use of the successful external language model (ELM) to integrate the abundant linguistic knowledge into the caption model. The ELM generates more semantically similar word proposals which extend the groundtruth words used for training to deal with the long-tailed problem. Experimental evaluations on thre e benchmarks: MSVD, MSR-VTT and VATEX show the proposed ORG-TRL system achieves state-of-the-art performance. Extensive ablation studies and visualizations illu strate the effectiveness of our system.
****

Texture and Shape Biased Two-Stream Networks for Clothing Classification and Attribute Recognition
Yuwei Zhang, Peng Zhang, Chun Yuan, Zhi Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13538-13547
Clothes category classification and attribute recognition have achieved distinguished success with the development of deep learning. People have found that landmark detection plays a positive role in these tasks. However, little research is committed to analyzing these tasks from the perspective of clothing attributes. In our work, we explore the usefulness of landmarks and find that landmarks can assist in extracting shape features; and using landmarks for joint learning can increase classification and recognition accuracy effectively. We also find that texture features have an impelling effect on these tasks and that the pre-trained ImageNet model has good performance in extracting texture features. To this end, we propose to use two streams to enhance the extraction of shape and texture, respectively. In particular, this paper proposes a simple implementation, Texture and Shape biased Fashion Networks (TS-FashionNet). Comprehensive and rich experiments demonstrate our discoveries and the effectiveness of our model. We improve the top-3 classification accuracy by 0.83% and improve the top-3 attribute recognition recall rate by 1.39% compared to the state-of-the-art models.
****
Combining Detection and Tracking for Human Pose Estimation in Videos
Manchen Wang, Joseph Tighe, Davide Modolo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11088-11096
We propose a novel top-down approach that tackles the problem of multi-person human pose estimation and tracking in videos. In contrast to existing top-down approaches, our method is not limited by the performance of its person detector and can predict the poses of person instances not localized. It achieves this capability by propagating known person locations forward and backward in time and searching for poses in those regions. Our approach consists of three components: (i) a Clip Tracking Network that performs body joint detection and tracking simultaneously on small video clips; (ii) a Video Tracking Pipeline that merges the fixed-length tracklets produced by the Clip Tracking Network to arbitrary length tracks; and (iii) a Spatial-Temporal Merging procedure that refines the joint locations based on spatial and temporal smoothing terms. Thanks to the precision of our Clip Tracking Network and our merging procedure, our approach produces very accurate joint predictions and can fix common mistakes on hard scenarios like heavily entangled people. Our approach achieves state-of-the-art results on both joint detection and tracking, on both the PoseTrack 2017 and 2018 datasets, and against all top-down and bottom-down approaches.
****
Few Sample Knowledge Distillation for Efficient Network Compression
Tianhong Li, Jianguo Li, Zhuang Liu, Changshui Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14639-14647
Deep neural network compression techniques such as pruning and weight tensor decomposition usually require fine-tuning to recover the prediction accuracy when the compression ratio is high. However, conventional fine-tuning suffers from the requirement of a large training set and the time-consuming training procedure. This paper proposes a novel solution for knowledge distillation from label-free few samples to realize both data efficiency and training/processing efficiency. We treat the original network as "teacher-net" and the compressed network as "student-net". A 1x1 convolution layer is added at the end of each layer block of the student-net, and we fit the block-level outputs of the student-net to the teacher-net by estimating the parameters of the added layers. We prove that the added layer can be merged without adding extra parameters and computation cost during inference. Experiments on multiple datasets and network architectures verify the method's effectiveness on student-nets obtained by various network pruning and weight decomposition methods. Our method can recover student-net's accuracy to the same level as conventional fine-tuning methods in minutes while using only 1% label-free data of the full training data.

****
Context R-CNN: Long Term Temporal Context for Per-Camera Object Detection
Sara Beery, Guanhang Wu, Vivek Rathod, Ronny Votel, Jonathan Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13075-13085
In static monitoring cameras, useful contextual information can stretch far beyond the few seconds typical video understanding models might see: subjects may exhibit similar behavior over multiple days, and background objects remain static. Due to power and storage constraints, sampling frequencies are low, often no faster than one frame per second, and sometimes are irregular due to the use of a motion trigger. In order to perform well in this setting, models must be robust to irregular sampling rates. In this paper we propose a method that leverages temporal context from the unlabeled frames of a novel camera to improve performance at that camera. Specifically, we propose an attention-based approach that allows our model, Context R-CNN, to index into a long term memory bank constructed on a per-camera basis and aggregate contextual features from other frames to boost object detection performance on the current frame. We apply Context R-CNN to two settings: (1) species detection using camera traps, and (2) vehicle detection in traffic cameras, showing in both settings that Context R-CNN leads to performance gains over strong baselines. Moreover, we show that increasing the contextual time horizon leads to improved results. When applied to camera trap data from the Snapshot Serengeti dataset, Context R-CNN with context from up to a month of images outperforms a single-frame baseline by 17.9% mAP, and outperforms S3D (a 3d convolution based baseline) by 11.2% mAP.
****
Temporal-Context Enhanced Detection of Heavily Occluded Pedestrians
Jialian Wu, Chunluan Zhou, Ming Yang, Qian Zhang, Yuan Li, Junsong Yuan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13430-13439
State-of-the-art pedestrian detectors have performed promisingly on non-occluded pedestrians, yet they are still confronted by heavy occlusions. Although many previous works have attempted to alleviate the pedestrian occlusion issue, most of them rest on still images. In this paper, we exploit the local temporal context of pedestrians in videos and propose a tube feature aggregation network (TFAN) aiming at enhancing pedestrian detectors against severe occlusions. Specifically, for an occluded pedestrian in the current frame, we iteratively search for its relevant counterparts along temporal axis to form a tube. Then, features from the tube are aggregated according to an adaptive weight to enhance the feature representations of the occluded pedestrian. Furthermore, we devise a temporally discriminative embedding module (TDEM) and a part-based relation module (PRM), respectively, which adapts our approach to better handle tube drifting and heavy occlusions. Extensive experiments are conducted on three datasets, Caltech, NightOwls and KAIST, showing that our proposed method is significantly effective for heavily occluded pedestrian detection. Moreover, we achieve the state-of-the-art performance on the Caltech and NightOwls datasets.
****
NMS by Representative Region: Towards Crowded Pedestrian Detection by Proposal Pairing
Xin Huang, Zheng Ge, Zequn Jie, Osamu Yoshie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10750-10759
Although significant progress has been made in pedestrian detection recently, pedestrian detection in crowded scenes is still challenging. The heavy occlusion between pedestrians imposes great challenges to the standard Non-Maximum Suppression (NMS). A relative low threshold of intersection over union (IoU) leads to missing highly overlapped pedestrians, while a higher one brings in plenty of false positives. To avoid such a dilemma, this paper proposes a novel Representative Region NMS (R2NMS) approach leveraging the less occluded visible parts, effectively removing the redundant boxes without bringing in many false positives. To acquire the visible parts, a novel Paired-Box Model (PBM) is proposed to simultaneously predict the full and visible boxes of a pedestrian. The full and visible

boxes constitute a pair serving as the sample unit of the model, thus guaranteeing a strong correspondence between the two boxes throughout the detection pipeline. Moreover, convenient feature integration of the two boxes is allowed for the better performance on both full and visible pedestrian detection tasks. Experiments on the challenging CrowdHuman and CityPersons benchmarks sufficiently validate the effectiveness of the proposed approach on pedestrian detection in the crowded situation.

****

PhraseCut: Language-Based Image Segmentation in the Wild

Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, Subhransu Maji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10216-10225

We consider the problem of segmenting image regions given a natural language phrase, and study it on a novel dataset of 77,262 images and 345,486 phrase-region pairs. Our dataset is collected on top of the Visual Genome dataset and uses the existing annotations to generate a challenging set of referring phrases for which the corresponding regions are manually annotated. Phrases in our dataset correspond to multiple regions and describe a large number of object and stuff categories as well as their attributes such as color, shape, parts, and relationships with other entities in the image. Our experiments show that the scale and diversity of concepts in our dataset poses significant challenges to the existing state-of-the-art. We systematically handle the long-tail nature of these concepts and present a modular approach to combine category, attribute, and relationship cues that outperforms existing approaches.

****

Learning User Representations for Open Vocabulary Image Hashtag Prediction

Thibaut Durand; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9769-9778

In this paper, we introduce an open vocabulary model for image hashtag prediction - the task of mapping an image to its accompanying hashtags. Recent work shows that to build an accurate hashtag prediction model, it is necessary to model the user because of the self-expression problem, in which similar image content may be labeled with different tags. To take into account the user behaviour, we propose a new model that extracts a representation of a user based on his/her image history. Our model allows to improve a user representation with new images or add a new user without retraining the model. Because new hashtags appear all the time on social networks, we design an open vocabulary model which can deal with new hashtags without retraining the model. Our model learns a cross-modal embedding between user conditional visual representations and hashtag word representations. Experiments on a subset of the YFCC100M dataset demonstrate the efficacy of our user representation in user conditional hashtag prediction and user retrieval. We further validate the open vocabulary prediction ability of our model.

****

PFCNN: Convolutional Neural Networks on 3D Surfaces Using Parallel Frames

Yuqi Yang, Shilin Liu, Hao Pan, Yang Liu, Xin Tong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13578-13587

Surface meshes are widely used shape representations and capture finer geometry data than point clouds or volumetric grids, but are challenging to apply CNNs directly due to their non-Euclidean structure. We use parallel frames on surface to define PFCNNs that enable effective feature learning on surface meshes by mimicking standard convolutions faithfully. In particular, the convolution of PFCNN not only maps local surface patches onto flat tangent planes, but also aligns the tangent planes such that they locally form a flat Euclidean structure, thus enabling recovery of standard convolutions. The alignment is achieved by the tool of locally flat connections borrowed from discrete differential geometry, which can be efficiently encoded and computed by parallel frame fields. In addition, the lack of canonical axis on surface is handled by sampling with the frame directions. Experiments show that for tasks including classification, segmentation and registration on deformable geometric domains, as well as semantic scene segmen

tation on rigid domains, PFCNNs achieve robust and superior performances without using sophisticated input features than state-of-the-art surface based CNNs.
****

Learning Weighted Submanifolds With Variational Autoencoders and Riemannian Variational Autoencoders
Nina Miolane, Susan Holmes; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14503-14511
Manifold-valued data naturally arises in medical imaging. In cognitive neuroscience for instance, brain connectomes base the analysis of coactivation patterns between different brain regions on the analysis of the correlations of their functional Magnetic Resonance Imaging (fMRI) time series - an object thus constrained by construction to belong to the manifold of symmetric positive definite matrices. One of the challenges that naturally arises in these studies consists in finding a lower-dimensional subspace for representing such manifold-valued and typically high-dimensional data. Traditional techniques, like principal component analysis, are ill-adapted to tackle non-Euclidean spaces and may fail to achieve a lower-dimensional representation of the data - thus potentially pointing to the absence of lower-dimensional representation of the data. However, these techniques are restricted in that: (i) they do not leverage the assumption that the connectomes belong on a pre-specified manifold, therefore discarding information; (ii) they can only fit a linear subspace to the data. In this paper, we are interested in variants to learn potentially highly curved submanifolds of manifold-valued data. Motivated by the brain connectomes example, we investigate a latent variable generative model, which has the added benefit of providing us with uncertainty estimates - a crucial quantity in the medical applications we are considering. While latent variable models have been proposed to learn linear and nonlinear spaces for Euclidean data, or geodesic subspaces for manifold data, no intrinsic latent variable model exists to learn non-geodesic subspaces for manifold data. This paper fills this gap and formulates a Riemannian variational autoencoder with an intrinsic generative model of manifold-valued data. We evaluate its performances on synthetic and real datasets, by introducing the formalism of weighted Riemannian submanifolds.
****

Learning Situational Driving
Eshed Ohn-Bar, Aditya Prakash, Aseem Behl, Kashyap Chitta, Andreas Geiger; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11296-11305
Human drivers have a remarkable ability to drive in diverse visual conditions and situations, e.g., from maneuvering in rainy, limited visibility conditions with no lane markings to turning in a busy intersection while yielding to pedestrians. In contrast, we find that state-of-the-art sensorimotor driving models struggle when encountering diverse settings with varying relationships between observation and action. To generalize when making decisions across diverse conditions, humans leverage multiple types of situation-specific reasoning and learning strategies. Motivated by this observation, we develop a framework for learning a situational driving policy that effectively captures reasoning under varying types of scenarios. Our key idea is to learn a mixture model with a set of policies that can capture multiple driving modes. We first optimize the mixture model through behavior cloning and show it to result in significant gains in terms of driving performance in diverse conditions. We then refine the model by directly optimizing for the driving task itself, i.e., supervised with the navigation task reward. Our method is more scalable than methods assuming access to privileged information, e.g., perception labels, as it only assumes demonstration and reward-based supervision. We achieve over 98% success rate on the CARLA driving benchmark as well as state-of-the-art performance on a newly introduced generalization benchmark.
****

Pose-Guided Visible Part Matching for Occluded Person ReID
Shang Gao, Jingya Wang, Huchuan Lu, Zimo Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11744-11752

Occluded person re-identification is a challenging task as the appearance varies substantially with various obstacles, especially in the crowd scenario. To address this issue, we propose a Pose-guided Visible Part Matching (PVPM) method that jointly learns the discriminative features with pose-guided attention and self-mines the part visibility in an end-to-end framework. Specifically, the proposed PVPM includes two key components: 1) pose-guided attention (PGA) method for part feature pooling that exploits more discriminative local features; 2) pose-guided visibility predictor (PVP) that estimates whether a part suffers the occlusion or not. As there are no ground truth training annotations for the occluded part, we turn to utilize the characteristic of part correspondence in positive pairs and self-mining the correspondence scores via graph matching. The generated correspondence scores are then utilized as pseudo-labels for visibility predictor (PVP). Experimental results on three reported occluded benchmarks show that the proposed method achieves competitive performance to state-of-the-art methods. The source codes are available at https://github.com/hh23333/PVPM

****

Online Knowledge Distillation via Collaborative Learning

Qiushan Guo, Xinjiang Wang, Yichao Wu, Zhipeng Yu, Ding Liang, Xiaolin Hu, Ping Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11020-11029

This work presents an efficient yet effective online Knowledge Distillation method via Collaborative Learning, termed KDCL, which is able to consistently improve the generalization ability of deep neural networks (DNNs) that have different learning capacities. Unlike existing two-stage knowledge distillation approaches that pre-train a DNN with large capacity as the "teacher" and then transfer the teacher's knowledge to another "student" DNN unidirectionally (i.e. one-way), KDCL treats all DNNs as "students" and collaboratively trains them in a single stage (knowledge is transferred among arbitrary students during collaborative training), enabling parallel computing, fast computations, and appealing generalization ability. Specifically, we carefully design multiple methods to generate soft target as supervisions by effectively ensembling predictions of students and distorting the input images. Extensive experiments show that KDCL consistently improves all the "students" on different datasets, including CIFAR-100 and ImageNet. For example, when trained together by using KDCL, ResNet-50 and MobileNetV2 achieve 78.2% and 74.0% top-1 accuracy on ImageNet, outperforming the original results by 1.4% and 2.0% respectively. We also verify that models pre-trained with KDCL transfer well to object detection and semantic segmentation on MS COCO dataset. For instance, the FPN detector is improved by 0.9% mAP.

****

Probabilistic Pixel-Adaptive Refinement Networks

Anne S. Wannenwetsch, Stefan Roth; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11642-11651

Encoder-decoder networks have found widespread use in various dense prediction tasks. However, the strong reduction of spatial resolution in the encoder leads to a loss of location information as well as boundary artifacts. To address this, image-adaptive post-processing methods have shown beneficial by leveraging the high-resolution input image(s) as guidance data. We extend such approaches by considering an important orthogonal source of information: the network's confidence in its own predictions. We introduce probabilistic pixel-adaptive convolutions (PPACs), which not only depend on image guidance data for filtering, but also respect the reliability of per-pixel predictions. As such, PPACs allow for image-adaptive smoothing and simultaneously propagating pixels of high confidence into less reliable regions, while respecting object boundaries. We demonstrate their utility in refinement networks for optical flow and semantic segmentation, where PPACs lead to a clear reduction in boundary artifacts. Moreover, our proposed refinement step is able to substantially improve the accuracy on various widely used benchmarks.

****

"Looking at the Right Stuff" - Guided Semantic-Gaze for Autonomous Driving

Anwesan Pal, Sayan Mondal, Henrik I. Christensen; Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11883-11892

In recent years, predicting driver's focus of attention has been a very active area of research in the autonomous driving community. Unfortunately, existing state-of-the-art techniques achieve this by relying only on human gaze information, thereby ignoring scene semantics. We propose a novel Semantics Augmented GazE (SAGE) detection approach that captures driving specific contextual information, in addition to the raw gaze. Such a combined attention mechanism serves as a powerful tool to focus on the relevant regions in an image frame in order to make driving both safe and efficient. Using this, we design a complete saliency prediction framework - SAGE-Net, which modifies the initial prediction from SAGE by taking into account vital aspects such as distance to objects (depth), ego vehicle speed, and pedestrian crossing intent. Exhaustive experiments conducted through four popular saliency algorithms show that on 49/56 (87.5%) cases - considering both the overall dataset and crucial driving scenarios, SAGE outperforms existing techniques without any additional computational overhead during the training process. The augmented dataset along with the relevant code are available as part of the supplementary material.

****

Social-STGCNN: A Social Spatio-Temporal Graph Convolutional Neural Network for Human Trajectory Prediction

Abduallah Mohamed, Kun Qian, Mohamed Elhoseiny, Christian Claudel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14424-14432

Better machine understanding of pedestrian behaviors enables faster progress in modeling interactions between agents such as autonomous vehicles and humans. Pedestrian trajectories are not only influenced by the pedestrian itself but also by interaction with surrounding objects. Previous methods modeled these interactions by using a variety of aggregation methods that integrate different learned pedestrians states. We propose the Social Spatio-Temporal Graph Convolutional Neural Network (Social-STGCNN), which substitutes the need of aggregation methods by modeling the interactions as a graph. Our results show an improvement over the state of art by 20% on the Final Displacement Error (FDE) and an improvement on the Average Displacement Error (ADE) with 8.5 times less parameters and up to 48 times faster inference speed than previously reported methods. In addition, our model is data efficient, and exceeds previous state of the art on the ADE metric with only 20% of the training data. We propose a kernel function to embed the social interactions between pedestrians within the adjacency matrix. Through qualitative analysis, we show that our model inherited social behaviors that can be expected between pedestrians trajectories. Code is available at https://github.com/abduallahmohamed/Social-STGCNN.

****

Efficient Neural Vision Systems Based on Convolutional Image Acquisition

Pedram Pad, Simon Narduzzi, Clement Kundig, Engin Turetken, Siavash A. Bigdeli, L. Andrea Dunbar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12285-12294

Despite the substantial progress made in deep learning in recent years, advanced approaches remain computationally intensive. The trade-off between accuracy and computation time and energy limits their use in real-time applications on low power and other resource-constrained systems. In this paper, we tackle this fundamental challenge by introducing a hybrid optical-digital implementation of a convolutional neural network (CNN) based on engineering of the point spread function (PSF) of an optical imaging system. This is done by coding an imaging aperture such that its PSF replicates a large convolution kernel of the first layer of a pre-trained CNN. As the convolution takes place in the optical domain, it has zero cost in terms of energy consumption and has zero latency independent of the kernel size. Experimental results on two datasets demonstrate that our approach yields more than two orders of magnitude reduction in the computational cost while achieving near-state-of-the-art accuracy, or equivalently, better accuracy at the same computational cost.

****
DAVD-Net: Deep Audio-Aided Video Decompression of Talking Heads
Xi Zhang, Xiaolin Wu, Xinliang Zhai, Xianye Ben, Chengjie Tu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12335-12344
Close-up talking heads are among the most common and salient object in video contents, such as face-to-face conversations in social media, teleconferences, news broadcasting, talk shows, etc. Due to the high sensitivity of human visual system to faces, compression distortions in talking heads videos are highly visible and annoying. To address this problem, we present a novel deep convolutional neural network (DCNN) method for very low bit rate video reconstruction of talking heads. The key innovation is a new DCNN architecture that can exploit the audio-video correlations to repair compression defects in the face region. We further improve reconstruction quality by embedding into our DCNN the encoder information of the video compression standards and introducing a constraining projection module in the network. Extensive experiments demonstrate that the proposed DCNN method outperforms the existing state-of-the-art methods on videos of talking heads.
****
Referring Image Segmentation via Cross-Modal Progressive Comprehension
Shaofei Huang, Tianrui Hui, Si Liu, Guanbin Li, Yunchao Wei, Jizhong Han, Luoqi Liu, Bo Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10488-10497
Referring image segmentation aims at segmenting the foreground masks of the entities that can well match the description given in the natural language expression. Previous approaches tackle this problem using implicit feature interaction and fusion between visual and linguistic modalities, but usually fail to explore informative words of the expression to well align features from the two modalities for accurately identifying the referred entity. In this paper, we propose a Cross-Modal Progressive Comprehension (CMPC) module and a Text-Guided Feature Exchange (TGFE) module to effectively address the challenging task. Concretely, the CMPC module first employs entity and attribute words to perceive all the related entities that might be considered by the expression. Then, the relational words are adopted to highlight the correct entity as well as suppress other irrelevant ones by multimodal graph reasoning. In addition to the CMPC module, we further leverage a simple yet effective TGFE module to integrate the reasoned multimodal features from different levels with the guidance of textual information. In this way, features from multi-levels could communicate with each other and be refined based on the textual context. We conduct extensive experiments on four popular referring segmentation benchmarks and achieve new state-of-the-art performances. Code is available at https://github.com/spyflying/CMPC-Refseg.
****
SAPIEN: A SimulAted Part-Based Interactive ENvironment
Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, Li Yi, Angel X. Chang, Leonidas J. Guibas, Hao Su; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11097-11107
Building home assistant robots has long been a goal for vision and robotics researchers. To achieve this task, a simulated environment with physically realistic simulation, sufficient articulated objects, and transferability to the real robot is indispensable. Existing environments achieve these requirements for robotics simulation with different levels of simplification and focus. We take one step further in constructing an environment that supports household tasks for training robot learning algorithm. Our work, SAPIEN, is a realistic and physics-rich simulated environment that hosts a large-scale set of articulated objects. SAPIEN enables various robotic vision and interaction tasks that require detailed part-level understanding.We evaluate state-of-the-art vision algorithms for part detection and motion attribute recognition as well as demonstrate robotic interaction tasks using heuristic approaches and reinforcement learning algorithms. We hope that SAPIEN will open research directions yet to be explored, including lear

ning cognition through interaction, part motion discovery, and construction of robotics-ready simulated game environment.
****

Appearance Shock Grammar for Fast Medial Axis Extraction From Real Images
Charles-Olivier Dufresne Camaro, Morteza Rezanejad, Stavros Tsogkas, Kaleem Siddiqi, Sven Dickinson; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14382-14391
We combine ideas from shock graph theory with more recent appearance-based methods for medial axis extraction from complex natural scenes, improving upon the present best unsupervised method, in terms of efficiency and performance. We make the following specific contributions: i) we extend the shock graph representation to the domain of real images, by generalizing the shock type definitions using local, appearance-based criteria; ii) we then use the rules of a Shock Grammar to guide our search for medial points, drastically reducing run time when compared to other methods, which exhaustively consider all points in the input image; iii) we remove the need for typical post-processing steps including thinning, non-maximum suppression, and grouping, by adhering to the Shock Grammar rules while deriving the medial axis solution; iv) finally, we raise some fundamental concerns with the evaluation scheme used in previous work and propose a more appropriate alternative for assessing the performance of medial axis extraction from scenes. Our experiments on the BMAX500 and SK-LARGE datasets demonstrate the effectiveness of our approach. We outperform the present state-of-the-art, excelling particularly in the high-precision regime, while running an order of magnitude faster and requiring no post-processing.
****

TransMatch: A Transfer-Learning Scheme for Semi-Supervised Few-Shot Learning
Zhongjie Yu, Lin Chen, Zhongwei Cheng, Jiebo Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12856-12864
The successful application of deep learning to many visual recognition tasks relies heavily on the availability of a large amount of labeled data which is usually expensive to obtain. The few-shot learning problem has attracted increasing attention from researchers for building a robust model upon only a few labeled samples. Most existing works tackle this problem under the meta-learning framework by mimicking the few-shot learning task with an episodic training strategy. In this paper, we propose a new transfer-learning framework for semi-supervised few-shot learning to fully utilize the auxiliary information from labeled base-class data and unlabeled novel-class data. The framework consists of three components: 1) pre-training a feature extractor on base-class data; 2) using the feature extractor to initialize the classifier weights for the novel classes; and 3) further updating the model with a semi-supervised learning method. Under the proposed framework, we develop a novel method for semi-supervised few-shot learning called TransMatch by instantiating the three components with imprinting and MixMatch. Extensive experiments on two popular benchmark datasets for few-shot learning, CUB-200-2011 and miniImageNet, demonstrate that our proposed method can effectively utilize the auxiliary information from labeled base-class data and unlabeled novel-class data to significantly improve the accuracy of few-shot learning task, and achieve new state-of-the-art results.
****

Solving Mixed-Modal Jigsaw Puzzle for Fine-Grained Sketch-Based Image Retrieval
Kaiyue Pang, Yongxin Yang, Timothy M. Hospedales, Tao Xiang, Yi-Zhe Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10347-10355
ImageNet pre-training has long been considered crucial by the fine-grained sketch-based image retrieval (FG-SBIR) community due to the lack of large sketch-photo paired datasets for FG-SBIR training. In this paper, we propose a self-supervised alternative for representation pre-training. Specifically, we consider the jigsaw puzzle game of recomposing images from shuffled parts. We identify two key facets of jigsaw task design that are required for effective FG-SBIR pre-training. The first is formulating the puzzle in a mixed-modality fashion. Second we s

how that framing the optimisation as permutation matrix inference via Sinkhorn i
terations is more effective than the common classifier formulation of Jigsaw sel
f-supervision. Experiments show that this self-supervised pre-training strategy
significantly outperforms the standard ImageNet-based pipeline across all four p
roduct-level FG-SBIR benchmarks. Interestingly it also leads to improved cross-c
ategory generalisation across both pre-train/fine-tune and fine-tune/testing sta
ges.
****

PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection
Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang,
Hongsheng Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pa
ttern Recognition (CVPR), 2020, pp. 10529-10538
We present a novel and high-performance 3D object detection framework, named Poi
ntVoxel-RCNN (PV-RCNN), for accurate 3D object detection from point clouds. Our
proposed method deeply integrates both 3D voxel Convolutional Neural Network (CN
N) and PointNet-based set abstraction to learn more discriminative point cloud f
eatures. It takes advantages of efficient learning and high-quality proposals of
the 3D voxel CNN and the flexible receptive fields of the PointNet-based networ
ks. Specifically, the proposed framework summarizes the 3D scene with a 3D voxel
CNN into a small set of keypoints via a novel voxel set abstraction module to s
ave follow-up computations and also to encode representative scene features. Giv
en the high-quality 3D proposals generated by the voxel CNN, the RoI-grid poolin
g is proposed to abstract proposal-specific features from the keypoints to the R
oI-grid points via keypoint set abstraction. Compared with conventional pooling
operations, the RoI-grid feature points encode much richer context information f
or accurately estimating object confidences and locations. Extensive experiments
on both the KITTI dataset and the Waymo Open dataset show that our proposed PV-
RCNN surpasses state-of-the-art 3D detection methods with remarkable margins.
****

A Real-Time Cross-Modality Correlation Filtering Method for Referring Expression
Comprehension
Yue Liao, Si Liu, Guanbin Li, Fei Wang, Yanjie Chen, Chen Qian, Bo Li; Pro
ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (
CVPR), 2020, pp. 10880-10889
Referring expression comprehension aims to localize the object instance describe
d by a natural language expression. Current referring expression methods have ac
hieved good performance. However, none of them is able to achieve real-time infe
rence without accuracy drop. The reason for the relatively slow inference speed
is that these methods artificially split the referring expression comprehension
into two sequential stages including proposal generation and proposal ranking. I
t does not exactly conform to the habit of human cognition. To this end, we prop
ose a novel Realtime Cross-modality Correlation Filtering method (RCCF). RCCF re
formulates the referring expression comprehension as a correlation filtering pro
cess. The expression is first mapped from the language domain to the visual doma
in and then treated as a template (kernel) to perform correlation filtering on t
he image feature map. The peak value in the correlation heatmap indicates the ce
nter points of the target box. In addition, RCCF also regresses a 2-D object siz
e and 2-D offset. The center point coordinates, object size and center point off
set together to form the target bounding box. Our method runs at 40 FPS while ac
hieving leading performance in RefClef, RefCOCO, RefCOCO+ and RefCOCOg benchmark
s. In the challenging RefClef dataset, our methods almost double the state-of-th
e-art performance (34.70% increased to 63.79%). We hope this work can arouse mor
e attention and studies to the new cross-modality correlation filtering framewor
k as well as the one-stage framework for referring expression comprehension.
****

Cross-Modal Cross-Domain Moment Alignment Network for Person Search
Ya Jing, Wei Wang, Liang Wang, Tieniu Tan; Proceedings of the IEEE/CVF Confer
ence on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10678-10686
Text-based person search has drawn increasing attention due to its wide applicat
ions in video surveillance. However, most of the existing models depend heavily

on paired image-text data, which is very expensive to acquire. Moreover, they always face huge performance drop when directly exploiting them to new domains. To overcome this problem, we make the first attempt to adapt the model to new target domains in the absence of pairwise labels, which combines the challenges from both cross-modal (text-based) person search and cross-domain person search. Specially, we propose a moment alignment network (MAN) to solve the cross-modal cross-domain person search task in this paper. The idea is to learn three effective moment alignments including domain alignment (DA), cross-modal alignment (CA) and exemplar alignment (EA), which together can learn domain-invariant and semantic aligned cross-modal representations to improve model generalization. Extensive experiments are conducted on CUHK Person Description dataset (CUHK-PEDES) and Richly Annotated Pedestrian dataset (RAP). Experimental results show that our proposed model achieves the state-of-the-art performances on five transfer tasks.
****

Smooth Shells: Multi-Scale Shape Registration With Functional Maps
Marvin Eisenberger, Zorah Lahner, Daniel Cremers; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12265-12274
We propose a novel 3D shape correspondence method based on the iterative alignment of so-called smooth shells. Smooth shells define a series of coarse-to-fine shape approximations designed to work well with multiscale algorithms. The main idea is to first align rough approximations of the geometry and then add more and more details to refine the correspondence. We fuse classical shape registration with Functional Maps by embedding the input shapes into an intrinsic-extrinsic product space. Moreover, we disambiguate intrinsic symmetries by applying a surrogate based Markov chain Monte Carlo initialization. Our method naturally handles various types of noise that commonly occur in real scans, like non-isometry or incompatible meshing. Finally, we demonstrate state-of-the-art quantitative results on several datasets and show that our pipeline produces smoother, more realistic results than other automatic matching methods in real world applications.
****

PnPNet: End-to-End Perception and Prediction With Tracking in the Loop
Ming Liang, Bin Yang, Wenyuan Zeng, Yun Chen, Rui Hu, Sergio Casas, Raquel Urtasun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11553-11562
We tackle the problem of joint perception and motion forecasting in the context of self-driving vehicles. Towards this goal we propose PnPNet, an end-to-end model that takes as input sequential sensor data, and outputs at each time step object tracks and their future trajectories. The key component is a novel tracking module that generates object tracks online from detections and exploits trajectory level features for motion forecasting. Specifically, the object tracks get updated at each time step by solving both the data association problem and the trajectory estimation problem. Importantly, the whole model is end-to-end trainable and benefits from joint optimization of all tasks. We validate PnPNet on two large-scale driving datasets, and show significant improvements over the state-of-the-art with better occlusion recovery and more accurate future prediction.
****

Exemplar Normalization for Learning Deep Representation
Ruimao Zhang, Zhanglin Peng, Lingyun Wu, Zhen Li, Ping Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12726-12735
Normalization techniques are important in different advanced neural networks and different tasks. This work investigates a novel dynamic learning-to-normalize (L2N) problem by proposing Exemplar Normalization (EN), which is able to learn different normalization methods for different convolutional layers and image samples of a deep network. EN significantly improves the flexibility of the recently proposed switchable normalization (SN), which solves a static L2N problem by linearly combining several normalizers in each normalization layer (the combination is the same for all samples). Instead of directly employing a multi-layer perceptron (MLP) to learn data-dependent parameters as conditional batch normalizatio

n (cBN) did, the internal architecture of EN is carefully designed to stabilize its optimization, leading to many appealing benefits. (1) EN enables different convolutional layers, image samples, categories, benchmarks, and tasks to use different normalization methods, shedding light on analyzing them in a holistic view. (2) EN is effective for various network architectures and tasks. (3) It could replace any normalization layers in a deep network and still produce stable model training. Extensive experiments demonstrate the effectiveness of EN in a wide spectrum of tasks including image recognition, noisy label learning, and semantic segmentation. For example, by replacing BN in the ordinary ResNet50, improvement produced by EN is 300% more than that of SN on both ImageNet and the noisy WebVision dataset. The codes and models will be released.

****

Graph-Structured Referring Expression Reasoning in the Wild

Sibei Yang, Guanbin Li, Yizhou Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9952-9961

Grounding referring expressions aims to locate in an image an object referred to by a natural language expression. The linguistic structure of a referring expression provides a layout of reasoning over the visual contents, and it is often crucial to align and jointly understand the image and the referring expression. In this paper, we propose a scene graph guided modular network (SGMN), which performs reasoning over a semantic graph and a scene graph with neural modules under the guidance of the linguistic structure of the expression. In particular, we model the image as a structured semantic graph, and parse the expression into a language scene graph. The language scene graph not only decodes the linguistic structure of the expression, but also has a consistent representation with the image semantic graph. In addition to exploring structured solutions to grounding referring expressions, we also propose Ref-Reasoning, a large-scale real-world dataset for structured referring expression reasoning. We automatically generate referring expressions over the scene graphs of images using diverse expression templates and functional programs. This dataset is equipped with real-world visual contents as well as semantically rich expressions with different reasoning layouts. Experimental results show that our SGMN not only significantly outperforms existing state-of-the-art algorithms on the new Ref-Reasoning dataset, but also surpasses state-of-the-art structured methods on commonly used benchmark datasets. It can also provide interpretable visual evidences of reasoning.

****

Feature-Metric Registration: A Fast Semi-Supervised Approach for Robust Point Cloud Registration Without Correspondences

Xiaoshui Huang, Guofeng Mei, Jian Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11366-11374

We present a fast feature-metric point cloud registration framework, which enforces the optimisation of registration by minimising a feature-metric projection error without correspondences. The advantage of the feature-metric projection error is robust to noise, outliers and density difference in contrast to the geometric projection error. Besides, minimising the feature-metric projection error does not need to search the correspondences so that the optimisation speed is fast. The principle behind the proposed method is that the feature difference is smallest if point clouds are aligned very well. We train the proposed method in a semi-supervised or unsupervised approach, which requires limited or no registration label data. Experiments demonstrate our method obtains higher accuracy and robustness than the state-of-the-art methods. Besides, experimental results show that the proposed method can handle significant noise and density difference, and solve both same-source and cross-source point cloud registration.

****

The Garden of Forking Paths: Towards Multi-Future Trajectory Prediction

Junwei Liang, Lu Jiang, Kevin Murphy, Ting Yu, Alexander Hauptmann; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10508-10518

This paper studies the problem of predicting the distribution over multiple possible future paths of people as they move through various visual scenes. We make

two main contributions. The first contribution is a new dataset, created in a realistic 3D simulator, which is based on real world trajectory data, and then extrapolated by human annotators to achieve different latent goals. This provides the first benchmark for quantitative evaluation of the models to predict multi-future trajectories. The second contribution is a new model to generate multiple plausible future trajectories, which contains novel designs of using multi-scale location encodings and convolutional RNNs over graphs. We refer to our model as Multiverse. We show that our model achieves the best results on our dataset, as well as on the real-world VIRAT/ActEV dataset (which just contains one possible future).
****

PolarMask: Single Shot Instance Segmentation With Polar Representation
Enze Xie, Peize Sun, Xiaoge Song, Wenhai Wang, Xuebo Liu, Ding Liang, Chunhua Shen, Ping Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12193-12202
In this paper, we introduce an anchor-box free and single shot instance segmentation method, which is conceptually simple, fully convolutional and can be used by easily embedding it into most off-the-shelf detection methods. Our method, termed PolarMask, formulates the instance segmentation problem as predicting contour of instance through instance center classification and dense distance regression in a polar coordinate. Moreover, we propose two effective approaches to deal with sampling high-quality center examples and optimization for dense distance regression, respectively, which can significantly improve the performance and simplify the training process. Without any bells and whistles, PolarMask achieves 32.9% in mask mAP with single-model and single-scale training/testing on the challenging COCO dataset. For the first time, we show that the complexity of instance segmentation, in terms of both design and computation complexity, can be the same as bounding box object detection and this much simpler and flexible instance segmentation framework can achieve competitive accuracy. We hope that the proposed PolarMask framework can serve as a fundamental and strong baseline for single shot instance segmentation task.
****

Towards Learning a Generic Agent for Vision-and-Language Navigation via Pre-Training
Weituo Hao, Chunyuan Li, Xiujun Li, Lawrence Carin, Jianfeng Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13137-13146
Learning to navigate in a visual environment following natural-language instructions is a challenging task, because the multimodal inputs to the agent are highly variable, and the training data on a new task is often limited. In this paper, we present the first pre-training and fine-tuning paradigm for vision-and-language navigation (VLN) tasks. By training on a large amount of image-text-action triplets in a self-supervised learning manner, the pre-trained model provides generic representations of visual environments and language instructions. It can be easily used as a drop-in for existing VLN frameworks, leading to the proposed agent PREVALENT. It learns more effectively in new tasks and generalizes better in a previously unseen environment. The performance is validated on three VLN tasks. On the Room-to-Room benchmark, our model improves the state-of-the-art from 47% to 51% on success rate weighted by path length. Further, the learned representation is transferable to other VLN tasks. On two recent tasks, vision-and-dialog navigation and "Help, Anna!", the proposed PREVALENT leads to significant improvement over existing methods, achieving a new state of the art.
****

Boosting Few-Shot Learning With Adaptive Margin Loss
Aoxue Li, Weiran Huang, Xu Lan, Jiashi Feng, Zhenguo Li, Liwei Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12576-12584
Few-shot learning (FSL) has attracted increasing attention in recent years but remains challenging, due to the intrinsic difficulty in learning to generalize from a few examples. This paper proposes an adaptive margin principle to improve t

he generalization ability of metric-based meta-learning approaches for few-shot learning problems. Specifically, we first develop a class-relevant additive margin loss, where semantic similarity between each pair of classes is considered to separate samples in the feature embedding space from similar classes. Further, we incorporate the semantic context among all classes in a sampled training task and develop a task-relevant additive margin loss to better distinguish samples from different classes. Our adaptive margin method can be easily extended to a more realistic generalized FSL setting. Extensive experiments demonstrate that the proposed method can boost the performance of current metric-based meta-learning approaches, under both the standard FSL and generalized FSL settings.
****

From Depth What Can You See? Depth Completion via Auxiliary Image Reconstruction
Kaiyue Lu, Nick Barnes, Saeed Anwar, Liang Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11306-11315
Depth completion recovers dense depth from sparse measurements, e.g., LiDAR. Existing depth-only methods use sparse depth as the only input. However, these methods may fail to recover semantics consistent boundaries, or small/thin objects due to 1) the sparse nature of depth points and 2) the lack of images to provide semantic cues. This paper continues this line of research and aims to overcome the above shortcomings. The unique design of our depth completion model is that it simultaneously outputs a reconstructed image and a dense depth map. Specifically, we formulate image reconstruction from sparse depth as an auxiliary task during training that is supervised by the unlabelled gray-scale images. During testing, our system accepts sparse depth as the only input, i.e., the image is not required. Our design allows the depth completion network to learn complementary image features that help to better understand object structures. The extra supervision incurred by image reconstruction is minimal, because no annotations other than the image are needed. We evaluate our method on the KITTI depth completion benchmark and show that depth completion can be significantly improved via the auxiliary supervision of image reconstruction. Our algorithm consistently outperforms depth-only methods and is also effective for indoor scenes like NYUv2.
****

PuppeteerGAN: Arbitrary Portrait Animation With Semantic-Aware Appearance Transformation
Zhuo Chen, Chaoyue Wang, Bo Yuan, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13518-13527
Portrait animation, which aims to animate a still portrait to life using poses extracted from target frames, is an important technique for many real-world entertainment applications. Although recent works have achieved highly realistic results on synthesizing or controlling human head images, the puppeteering of arbitrary portraits is still confronted by the following challenges: 1) identity/personality mismatch; 2) training data/domain limitations; and 3) low-efficiency in training/fine-tuning. In this paper, we devised a novel two-stage framework called PuppeteerGAN for solving these challenges. Specifically, we first learn identity-preserved semantic segmentation animation which executes pose retargeting between any portraits. As a general representation, the semantic segmentation results could be adapted to different datasets, environmental conditions or appearance domains. Furthermore, the synthesized semantic segmentation is filled with the appearance of the source portrait. To this end, an appearance transformation network is presented to produce fidelity output by jointly considering the wrapping of semantic features and conditional generation. After training, the two networks can directly perform end-to-end inference on unseen subjects without any retraining or fine-tuning. Extensive experiments on cross-identity/domain/resolution situations demonstrate the superiority of the proposed PuppetterGAN over existing portrait animation methods in both generation quality and inference speed.
****

Active Speakers in Context
Juan Leon Alcazar, Fabian Caba, Long Mai, Federico Perazzi, Joon-Young Lee,

Pablo Arbelaez, Bernard Ghanem; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12465-12474
Current methods for active speaker detection focus on modeling audiovisual information from a single speaker. This strategy can be adequate for addressing single-speaker scenarios, but it prevents accurate detection when the task is to identify who of many candidate speakers are talking. This paper introduces the Active Speaker Context, a novel representation that models relationships between multiple speakers over long time horizons. Our new model learns pairwise and temporal relations from a structured ensemble of audiovisual observations. Our experiments show that a structured feature ensemble already benefits active speaker detection performance. We also find that the proposed Active Speaker Context improves the state-of-the-art on the AVA-ActiveSpeaker dataset achieving an mAP of 87.1%. Moreover, ablation studies verify that this result is a direct consequence of our long-term multi-speaker analysis.

****

3DSSD: Point-Based 3D Single Stage Object Detector
Zetong Yang, Yanan Sun, Shu Liu, Jiaya Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11040-11048
Prevalence of voxel-based 3D single-stage detectors contrast with underexplored point-based methods. In this paper, we present a lightweight point-based 3D single stage object detector 3DSSD to achieve decent balance of accuracy and efficiency. In this paradigm, all upsampling layers and the refinement stage, which are indispensable in all existing point-based methods, are abandoned. We instead propose a fusion sampling strategy in downsampling process to make detection on less representative points feasible. A delicate box prediction network, including a candidate generation layer and an anchor-free regression head with a 3D center-ness assignment strategy, is developed to meet the demand of high accuracy and speed. Our 3DSSD paradigm is an elegant single-stage anchor-free one. We evaluate it on widely used KITTI dataset and more challenging nuScenes dataset. Our method outperforms all state-of-the-art voxel-based single-stage methods by a large margin, and even yields comparable performance with two-stage point-based methods, with amazing inference speed of 25+ FPS, 2x faster than former state-of-the-art point-based methods.

****

Learning to Learn Cropping Models for Different Aspect Ratio Requirements
Debang Li, Junge Zhang, Kaiqi Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12685-12694
Image cropping aims at improving the framing of an image by removing its extraneous outer areas, which is widely used in the photography and printing industry. In some cases, the aspect ratio of cropping results is specified depending on some conditions. In this paper, we propose a meta-learning (learning to learn) based aspect ratio specified image cropping method called Mars, which can generate cropping results of different expected aspect ratios. In the proposed method, a base model and two meta-learners are obtained during the training stage. Given an aspect ratio in the test stage, a new model with new parameters can be generated from the base model. Specifically, the two meta-learners predict the parameters of the base model based on the given aspect ratio. The learning process of the proposed method is learning how to learn cropping models for different aspect ratio requirements, which is a typical meta-learning process. In the experiments, the proposed method is evaluated on three datasets and outperforms most state-of-the-art methods in terms of accuracy and speed. In addition, both the intermediate and final results show that the proposed model can predict different cropping windows for an image depending on different aspect ratio requirements.

****

nuScenes: A Multimodal Dataset for Autonomous Driving
Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, Oscar Beijbom; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11621-11631
Robust detection and tracking of objects is crucial for the deployment of autono

mous vehicle technology. Image based benchmark datasets have driven development in computer vision tasks such as object detection, tracking and segmentation of agents in the environment. Most autonomous vehicles, however, carry a combination of cameras and range sensors such as lidar and radar. As machine learning based methods for detection and tracking become more prevalent, there is a need to train and evaluate such methods on datasets containing range sensor data along with images. In this work we present nuTonomy scenes (nuScenes), the first dataset to carry the full autonomous vehicle sensor suite: 6 cameras, 5 radars and 1 lidar, all with full 360 degree field of view. nuScenes comprises 1000 scenes, each 20s long and fully annotated with 3D bounding boxes for 23 classes and 8 attributes. It has 7x as many annotations and 100x as many images as the pioneering KITTI dataset. We define novel 3D detection and tracking metrics. We also provide careful dataset analysis as well as baselines for lidar and image based detection and tracking. Data, development kit and more information are available online.

****

Learning Visual Emotion Representations From Web Data

Zijun Wei, Jianming Zhang, Zhe Lin, Joon-Young Lee, Niranjan Balasubramanian, Minh Hoai, Dimitris Samaras; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13106-13115

We present a scalable approach for learning powerful visual features for emotion recognition. A critical bottleneck in emotion recognition is the lack of large scale datasets that can be used for learning visual emotion features. To this end, we curate a webly derived large scale dataset, StockEmotion, which has more than a million images. StockEmotion uses 690 emotion related tags as labels giving us a fine-grained and diverse set of emotion labels, circumventing the difficulty in manually obtaining emotion annotations. We use this dataset to train a feature extraction network, EmotionNet, which we further regularize using joint text and visual embedding and text distillation. Our experimental results establish that EmotionNet trained on the StockEmotion dataset outperforms SOTA models on four different visual emotion tasks. An aded benefit of our joint embedding training approach is that EmotionNet achieves competitive zero-shot recognition performance against fully supervised baselines on a challenging visual emotion dataset, EMOTIC, which further highlights the generalizability of the learned emotion features.

****

Fine-Grained Video-Text Retrieval With Hierarchical Graph Reasoning

Shizhe Chen, Yida Zhao, Qin Jin, Qi Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10638-10647

Cross-modal retrieval between videos and texts has attracted growing attentions due to the rapid emergence of videos on the web. The current dominant approach is to learn a joint embedding space to measure cross-modal similarities. However, simple embeddings are insufficient to represent complicated visual and textual details, such as scenes, objects, actions and their compositions. To improve fine-grained video-text retrieval, we propose a Hierarchical Graph Reasoning (HGR) model, which decomposes video-text matching into global-to-local levels. The model disentangles text into a hierarchical semantic graph including three levels of events, actions, entities, and generates hierarchical textual embeddings via attention-based graph reasoning. Different levels of texts can guide the learning of diverse and hierarchical video representations for cross-modal matching to capture both global and local details. Experimental results on three video-text datasets demonstrate the advantages of our model. Such hierarchical decomposition also enables better generalization across datasets and improves the ability to distinguish fine-grained semantic differences. Code will be released at https://github.com/cshizhe/hgr_v2t.

****

Generative-Discriminative Feature Representations for Open-Set Recognition

Pramuditha Perera, Vlad I. Morariu, Rajiv Jain, Varun Manjunatha, Curtis Wigington, Vicente Ordonez, Vishal M. Patel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11814-11823

We address the problem of open-set recognition, where the goal is to determine i
f a given sample belongs to one of the classes used for training a model (known
classes). The main challenge in open-set recognition is to disentangle open-set
samples that produce high class activations from known-set samples. We propose t
wo techniques to force class activations of open-set samples to be low. First, w
e train a generative model for all known classes and then augment the input with
 the representation obtained from the generative model to learn a classifier. Th
is network learns to associate high classification probabilities both when image
 content is from the correct class as well as when the input and the reconstruct
ed image are consistent with each other. Second, we use self-supervision to forc
e the network to learn more informative featues when assigning class scores to i
mprove separation of classes from each other and from open-set samples. We evalu
ate the performance of the proposed method with recent open-set recognition work
s across three datasets, where we obtain state-of-the-art results.
****

RandLA-Net: Efficient Semantic Segmentation of Large-Scale Point Clouds
Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Ni
ki Trigoni, Andrew Markham; Proceedings of the IEEE/CVF Conference on Computer
Vision and Pattern Recognition (CVPR), 2020, pp. 11108-11117
We study the problem of efficient semantic segmentation for large-scale 3D point
 clouds. By relying on expensive sampling techniques or computationally heavy pr
e/post-processing steps, most existing approaches are only able to be trained an
d operate over small-scale point clouds. In this paper, we introduce RandLA-Net,
 an efficient and lightweight neural architecture to directly infer per-point se
mantics for large-scale point clouds. The key to our approach is to use random p
oint sampling instead of more complex point selection approaches. Although remar
kably computation and memory efficient, random sampling can discard key features
 by chance. To overcome this, we introduce a novel local feature aggregation mod
ule to progressively increase the receptive field for each 3D point, thereby eff
ectively preserving geometric details. Extensive experiments show that our RandL
A-Net can process 1 million points in a single pass with up to 200x faster than
existing approaches. Moreover, our RandLA-Net clearly surpasses state-of-the-art
 approaches for semantic segmentation on two large-scale benchmarks Semantic3D a
nd SemanticKITTI.
****

Learning to Structure an Image With Few Colors
Yunzhong Hou, Liang Zheng, Stephen Gould; Proceedings of the IEEE/CVF Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10116-10125
Color and structure are the two pillars that construct an image. Usually, the st
ructure is well expressed through a rich spectrum of colors, allowing objects in
 an image to be recognized by neural networks. However, under extreme limitation
s of color space, the structure tends to vanish, and thus a neural network might
 fail to understand the image. Interested in exploring this interplay between co
lor and structure, we study the scientific problem of identifying and preserving
 the most informative image structures while constraining the color space to jus
t a few bits, such that the resulting image can be recognized with possibly high
 accuracy. To this end, we propose a color quantization network, ColorCNN, which
 learns to structure the images from the classification loss in an end-to-end ma
nner. Given a color space size, ColorCNN quantizes colors in the original image
by generating a color index map and an RGB color palette. Then, this color-quant
ized image is fed to a pre-trained task network to evaluate its performance. In
our experiment, with only a 1-bit color space (i.e., two colors), the proposed n
etwork achieves 82.1% top-1 accuracy on the CIFAR10 dataset, outperforming tradi
tional color quantization methods by a large margin. For applications, when enco
ded with PNG, the proposed color quantization shows superiority over other image
 compression methods in the extremely low bit-rate regime. The code is available
 at https://github.com/hou-yz/color_distillation.
****

Discriminative Multi-Modality Speech Recognition
Bo Xu, Cheng Lu, Yandong Guo, Jacob Wang; Proceedings of the IEEE/CVF Confere

nce on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14433-14442
Vision is often used as a complementary modality for audio speech recognition (ASR), especially in the noisy environment where performance of solo audio modality significantly deteriorates. After combining visual modality, ASR is upgraded to the multi-modality speech recognition (MSR). In this paper, we propose a two-stage speech recognition model. In the first stage, the target voice is separated from background noises with help from the corresponding visual information of lip movements, making the model 'listen' clearly. At the second stage, the audio modality combines visual modality again to better understand the speech by a MSR sub-network, further improving the recognition rate. There are some other key contributions: we introduce a pseudo-3D residual convolution (P3D)-based visual front-end to extract more discriminative features; we upgrade the temporal convolution block from 1D ResNet with the temporal convolutional network (TCN), which is more suitable for the temporal tasks; the MSR sub-network is built on the top of Element-wise-Attention Gated Recurrent Unit (EleAtt-GRU), which is more effective than Transformer in long sequences. We conducted extensive experiments on the LRS3-TED and the LRW datasets. Our two-stage model (audio enhanced multi-modality speech recognition, AE-MSR) consistently achieves the state-of-the-art performance by a significant margin, which demonstrates the necessity and effectiveness of AE-MSR.
****
Improving Convolutional Networks With Self-Calibrated Convolutions
Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Changhu Wang, Jiashi Feng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10096-10105
Recent advances on CNNs are mostly devoted to designing more complex architectures to enhance their representation learning capacity. In this paper, we consider how to improve the basic convolutional feature transformation process of CNNs without tuning the model architectures. To this end, we present a novel self-calibrated convolutions that explicitly expand fields-of-view of each convolutional layers through internal communications and hence enrich the output features. In particular, unlike the standard convolutions that fuse spatial and channel-wise information using small kernels (e.g., 3x3), self-calibrated convolutions adaptively build long-range spatial and inter-channel dependencies around each spatial location through a novel self-calibration operation. Thus, it can help CNNs generate more discriminative representations by explicitly incorporating richer information. Our self-calibrated convolution design is simple and generic, and can be easily applied to augment standard convolutional layers without introducing extra parameters and complexity. Extensive experiments demonstrate that when applying self-calibrated convolutions into different backbones, our networks can significantly improve the baseline models in a variety of vision tasks, including image recognition, object detection, instance segmentation, and keypoint detection, with no need to change the network architectures. We hope this work could provide a promising way for future research in designing novel convolutional feature transformations for improving convolutional networks. Code is available on the project page.
****
CenterMask: Real-Time Anchor-Free Instance Segmentation
Youngwan Lee, Jongyoul Park; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13906-13915
We propose a simple yet efficient anchor-free instance segmentation, called CenterMask, that adds a novel spatial attention-guided mask (SAG-Mask) branch to anchor-free one stage object detector (FCOS) in the same vein with Mask R-CNN. Plugged into the FCOS object detector, the SAG-Mask branch predicts a segmentation mask on each box with the spatial attention map that helps to focus on informative pixels and suppress noise. We also present an improved backbone networks, VoVNetV2, with two effective strategies: (1) residual connection for alleviating the optimization problem of larger VoVNet [??] and (2) effective Squeeze-Excitation (eSE) dealing with the channel information loss problem of original SE. With SAG-Mask and VoVNetV2, we deign CenterMask and CenterMask-Lite that are targeted t

o large and small models, respectively. Using the same ResNet-101-FPN backbone, CenterMask achieves 38.3%, surpassing all previous state-of-the-art methods while at a much faster speed. CenterMask-Lite also outperforms the state-of-the-art by large margins at over 35fps on Titan Xp. We hope that CenterMask and VoVNetV2 can serve as a solid baseline of real-time instance segmentation and backbone network for various vision tasks, respectively. The Code is available at https://github.com/youngwanLEE/CenterMask.
****

Where Does It Exist: Spatio-Temporal Video Grounding for Multi-Form Sentences
Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, Lianli Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10668-10677
In this paper, we consider a novel task, Spatio-Temporal Video Grounding for Multi-Form Sentences (STVG). Given an untrimmed video and a declarative/interrogative sentence depicting an object, STVG aims to localize the spatio-temporal tube of the queried object. STVG has two challenging settings: (1) We need to localize spatio-temporal object tubes from untrimmed videos, where the object may only exist in a very small segment of the video; (2) We deal with multi-form sentences, including the declarative sentences with explicit objects and interrogative sentences with unknown objects. Existing methods cannot tackle the STVG task due to the ineffective tube pre-generation and the lack of object relationship modeling. Thus, we then propose a novel Spatio-Temporal Graph Reasoning Network (STGRN) for this task. First, we build a spatio-temporal region graph to capture the region relationships with temporal object dynamics, which involves the implicit and explicit spatial subgraphs in each frame and the temporal dynamic subgraph across frames. We then incorporate textual clues into the graph and develop the multi-step cross-modal graph reasoning. Next, we introduce a spatio-temporal localizer with a dynamic selection method to directly retrieve the spatio-temporal tubes without tube pre-generation. Moreover, we contribute a large-scale video grounding dataset VidSTG based on video relation dataset VidOR. The extensive experiments demonstrate the effectiveness of our method.
****

Autolabeling 3D Objects With Differentiable Rendering of SDF Shape Priors
Sergey Zakharov, Wadim Kehl, Arjun Bhargava, Adrien Gaidon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12224-12233
We present an automatic annotation pipeline to recover 9D cuboids and 3D shapes from pre-trained off-the-shelf 2D detectors and sparse LIDAR data. Our autolabeling method solves an ill-posed inverse problem by considering learned shape priors and optimizing geometric and physical parameters. To address this challenging problem, we apply a novel differentiable shape renderer to signed distance fields (SDF), leveraged together with normalized object coordinate spaces (NOCS). Initially trained on synthetic data to predict shape and coordinates, our method uses these predictions for projective and geometric alignment over real samples. Moreover, we also propose a curriculum learning strategy, iteratively retraining on samples of increasing difficulty in subsequent self-improving annotation rounds. Our experiments on the KITTI3D dataset show that we can recover a substantial amount of accurate cuboids, and that these autolabels can be used to train 3D vehicle detectors with state-of-the-art results.
****

Adaptive Fractional Dilated Convolution Network for Image Aesthetics Assessment
Qiuyu Chen, Wei Zhang, Ning Zhou, Peng Lei, Yi Xu, Yu Zheng, Jianping Fan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14114-14123
To leverage deep learning for image aesthetics assessment, one critical but unsolved issue is how to seamlessly incorporate the information of image aspect ratios to learn more robust models. In this paper, an adaptive fractional dilated convolution (AFDC), which is aspect-ratio-embedded, composition-preserving and parameter-free, is developed to tackle this issue natively in convolutional kernel level. Specifically, the fractional dilated kernel is adaptively constructed acc

ording to the image aspect ratios, where the interpolation of nearest two integer dilated kernels are used to cope with the misalignment of fractional sampling. Moreover, we provide a concise formulation for mini-batch training and utilize a grouping strategy to reduce computational overhead. As a result, it can be easily implemented by common deep learning libraries and plugged into popular CNN architectures in a computation-efficient manner. Our experimental results demonstrate that our proposed method achieves state-of-the-art performance on image aesthetics assessment over the AVA dataset.
****