Advisable Learning for Self-Driving Vehicles by Internalizing Observation-to-Action Rules
Jinkyu Kim, Suhong Moon, Anna Rohrbach, Trevor Darrell, John Canny; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9661-9670
Humans learn to drive through both practice and theory, e.g. by studying the rules, while most self-driving systems are limited to the former. Being able to incorporate human knowledge of typical causal driving behaviour should benefit autonomous systems. We propose a new approach that learns vehicle control with the help of human advice. Specifically, our system learns to summarize its visual observations in natural language, predict an appropriate action response (e.g. "I see a pedestrian crossing, so I stop"), and predict the controls, accordingly. Moreover, to enhance interpretability of our system, we introduce a fine-grained attention mechanism which relies on semantic segmentation and object-centric RoI pooling. We show that our approach of training the autonomous system with human advice, grounded in a rich semantic representation, matches or outperforms prior work in terms of control prediction and explanation generation. Our approach also results in more interpretable visual explanations by visualizing object-centric attention maps. Code is available at https://github.com/JinkyuKimUCB/advisable-driving.
****
Lightweight Multi-View 3D Pose Estimation Through Camera-Disentangled Representation
Edoardo Remelli, Shangchen Han, Sina Honari, Pascal Fua, Robert Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6040-6049
We present a lightweight solution to recover 3D pose from multi-view images captured with spatially calibrated cameras. Building upon recent advances in interpretable representation learning, we exploit 3D geometry to fuse input images into a unified latent representation of pose, which is disentangled from camera view-points. This allows us to reason effectively about 3D pose across different views without using compute-intensive volumetric grids. Our architecture then conditions the learned representation on camera projection operators to produce accurate per-view 2d detections, that can be simply lifted to 3D via a differentiable Direct Linear Transform (DLT) layer. In order to do it efficiently, we propose a novel implementation of DLT that is orders of magnitude faster on GPU architectures than standard SVD-based triangulation methods. We evaluate our approach on two large-scale human pose datasets (H36M and Total Capture): our method outperforms or performs comparably to the state-of-the-art volumetric methods, while, unlike them, yielding real-time performance.
****
Robust Design of Deep Neural Networks Against Adversarial Attacks Based on Lyapunov Theory
Arash Rahnama, Andre T. Nguyen, Edward Raff; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8178-8187
Deep neural networks (DNNs) are vulnerable to subtle adversarial perturbations applied to the input. These adversarial perturbations, though imperceptible, can easily mislead the DNN. In this work, we take a control theoretic approach to the problem of robustness in DNNs. We treat each individual layer of the DNN as a nonlinear system and use Lyapunov theory to prove stability and robustness locally. We then proceed to prove stability and robustness globally for the entire DNN. We develop empirically tight bounds on the response of the output layer, or any hidden layer, to adversarial perturbations added to the input, or the input of hidden layers. Recent works have proposed spectral norm regularization as a solution for improving robustness against l2 adversarial attacks. Our results give new insights into how spectral norm regularization can mitigate the adversarial effects. Finally, we evaluate the power of our approach on a variety of data sets and network architectures and against some of the well-known adversarial attacks.
****

Cross-Modal Deep Face Normals With Deactivable Skip Connections
Victoria Fernandez Abrevaya, Adnane Boukhayma, Philip H.S. Torr, Edmond Boyer ; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4979-4989
We present an approach for estimating surface normals from in-the-wild color images of faces. While data-driven strategies have been proposed for single face images, limited available ground truth data makes this problem difficult. To alleviate this issue, we propose a method that can leverage all available image and normal data, whether paired or not, thanks to a novel cross-modal learning architecture. In particular, we enable additional training with single modality data, either color or normal, by using two encoder-decoder networks with a shared latent space. The proposed architecture also enables face details to be transferred between the image and normal domains, given paired data, through skip connections between the image encoder and normal decoder. Core to our approach is a novel module that we call deactivable skip connections, which allows integrating both the auto-encoded and image-to-normal branches within the same architecture that can be trained end-to-end. This allows learning of a rich latent space that can accurately capture the normal information. We compare against state-of-the-art methods and show that our approach can achieve significant improvements, both quantitative and qualitative, with natural face images.
****
Progressive Adversarial Networks for Fine-Grained Domain Adaptation
Sinan Wang, Xinyang Chen, Yunbo Wang, Mingsheng Long, Jianmin Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9213-9222
Fine-grained visual categorization has long been considered as an important problem, however, its real application is still restricted, since precisely annotating a large fine-grained image dataset is a laborious task and requires expert-level human knowledge. A solution to this problem is applying domain adaptation approaches to fine-grained scenarios, where the key idea is to discover the commonality between existing fine-grained image datasets and massive unlabeled data in the wild. The main technical bottleneck lies in that the large inter-domain variation will deteriorate the subtle boundaries of small inter-class variation during domain alignment. This paper presents the Progressive Adversarial Networks (PAN) to align fine-grained categories across domains with a curriculum-based adversarial learning framework. In particular, throughout the learning process, domain adaptation is carried out through all multi-grained features, progressively exploiting the label hierarchy from coarse to fine. The progressive learning is applied upon both category classification and domain alignment, boosting both the discriminability and the transferability of the fine-grained features. Our method is evaluated on three benchmarks, two of which are proposed by us, and it outperforms the state-of-the-art domain adaptation methods.
****
ActBERT: Learning Global-Local Video-Text Representations
Linchao Zhu, Yi Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8746-8755
In this paper, we introduce ActBERT for self-supervised learning of joint video-text representations from unlabeled data. First, we leverage global action information to catalyze the mutual interactions between linguistic texts and local regional objects. It uncovers global and local visual clues from paired video sequences and text descriptions for detailed visual and text relation modeling. Second, we introduce an ENtangled Transformer block (ENT) to encode three sources of information, i.e., global actions, local regional objects, and linguistic descriptions. Global-local correspondences are discovered via judicious clues extraction from contextual information. It enforces the joint videotext representation to be aware of fine-grained objects as well as global human intention. We validate the generalization capability of ActBERT on downstream video-and language tasks, i.e., text-video clip retrieval, video captioning, video question answering, action segmentation, and action step localization. ActBERT significantly outperform the state-of-the-arts, demonstrating its superiority in video-text represen

tation learning.
****
Towards Visually Explaining Variational Autoencoders
Wenqian Liu, Runze Li, Meng Zheng, Srikrishna Karanam, Ziyan Wu, Bir Bhanu, Richard J. Radke, Octavia Camps; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8642-8651
Recent advances in Convolutional Neural Network (CNN) model interpretability have led to impressive progress in visualizing and understanding model predictions. In particular, gradient-based visual attention methods have driven much recent effort in using visual attention maps as a means for visual explanations. A key problem, however, is these methods are designed for classification and categorization tasks, and their extension to explaining generative models, e.g., variational autoencoders (VAE) is not trivial. In this work, we take a step towards bridging this crucial gap, proposing the first technique to visually explain VAEs by means of gradient-based attention. We present methods to generate visual attention from the learned latent space, and also demonstrate such attention explanations serve more than just explaining VAE predictions. We show how these attention maps can be used to localize anomalies in images, demonstrating state-of-the-art performance on the MVTec-AD dataset. We also show how they can be infused into model training, helping bootstrap the VAE into learning improved latent space disentanglement, demonstrated on the Dsprites dataset.
****
CenterMask: Single Shot Instance Segmentation With Point Representation
Yuqing Wang, Zhaoliang Xu, Hao Shen, Baoshan Cheng, Lirong Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9313-9321
In this paper, we propose a single-shot instance segmentation method, which is simple, fast and accurate. There are two main challenges for one-stage instance segmentation: object instances differentiation and pixel-wise feature alignment. Accordingly, we decompose the instance segmentation into two parallel subtasks: Local Shape prediction that separates instances even in overlapping conditions, and Global Saliency generation that segments the whole image in a pixel-to-pixel manner. The outputs of the two branches are assembled to form the final instance masks. To realize that, the local shape information is adopted from the representation of object center points. Totally trained from scratch and without any bells and whistles, the proposed CenterMask achieves 34.5 mask AP with a speed of 12.3 fps, using a single-model with single-scale training/testing on the challenging COCO dataset. The accuracy is higher than all other one-stage instance segmentation methods except the 5 times slower TensorMask, which shows the effectiveness of CenterMask. Besides, our method can be easily embedded to other one-stage object detectors such as FCOS and performs well, showing the generation of CenterMask.
****
BidNet: Binocular Image Dehazing Without Explicit Disparity Estimation
Yanwei Pang, Jing Nie, Jin Xie, Jungong Han, Xuelong Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5931-5940
Heavy haze results in severe image degradation and thus hampers the performance of visual perception, object detection, etc. On the assumption that dehazed binocular images are superior to the hazy ones for stereo vision tasks such as 3D object detection and according to the fact that image haze is a function of depth, this paper proposes a Binocular image dehazing Network (BidNet) aiming at dehazing both the left and right images of binocular images within the deep learning framework. Existing binocular dehazing methods rely on simultaneously dehazing and estimating disparity, whereas BidNet does not need to explicitly perform time-consuming and well-known challenging disparity estimation. Note that a small error in disparity gives rise to a large variation in depth and in estimation of haze-free image. The relationship and correlation between binocular images are explored and encoded by the proposed Stereo Transformation Module (STM). Jointly dehazing binocular image pairs is mutually beneficial, which is better than only

dehazing left images. We extend the Foggy Cityscapes dataset to a Stereo Foggy C ityscapes dataset with binocular foggy image pairs. Experimental results demonst rate that BidNet significantly outperforms state-of-the-art dehazing methods in both subjective and objective assessments.
****

Unsupervised Learning From Video With Deep Neural Embeddings
Chengxu Zhuang, Tianwei She, Alex Andonian, Max Sobol Mark, Daniel Yamins; P roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9563-9572
Because of the rich dynamical structure of videos andtheir ubiquity in everyday life, it is a natural idea that video data could serve as a powerful unsupervise d learning signal for visual representations. However, instantiating this idea, especially at large scale, has remained a significant artificial intelligence ch allenge. Here we present the Video Instance Embedding (VIE) framework, which tra ins deep nonlinear embeddings on video sequence inputs. By learning embedding di mensions that identify and group similar videos together, while pushing inherent ly different videos apart in the embedding space, VIE captures the strong statis tical structure inherent in videos, without the need for external annotation lab els. We find that, when trained on a large-scale video dataset, VIE yields power ful representations both for action recognition and single-frame object categori zation, showing substantially improving on the state of the art wherever direct comparisons are possible. We show that a two-pathway model with both static and dynamic processingpathways is optimal, provide analyses indicating how the model works, and perform ablation studies showing the importance of key architecture and loss function choices. Our results suggest that deep neural embeddings are a promising approach to unsupervised video learning fora wide variety of task dom ains.
****

One-Shot Domain Adaptation for Face Generation
Chao Yang, Ser-Nam Lim; Proceedings of the IEEE/CVF Conference on Computer Visi on and Pattern Recognition (CVPR), 2020, pp. 5921-5930
In this paper, we propose a framework capable of generating face images that fal l into the same distribution as that of a given one-shot example. We leverage a pre-trained StyleGAN model that already learned the generic face distribution. G iven the one-shot target, we develop an iterative optimization scheme that rapid ly adapts the weights of the model to shift the output's high-level distribution to the target's. To generate images of the same distribution, we introduce a st yle-mixing technique that transfers the low-level statistics from the target to faces randomly generated with the model. With that, we are able to generate an u nlimited number of faces that inherit from the distribution of both generic huma n faces and the one-shot example. The newly generated faces can serve as augment ed training data for other downstream tasks. Such setting is appealing as it req uires labeling very few, or even one example, in the target domain, which is oft en the case of real-world face manipulations that result from a variety of unkno wn and unique distributions, each with extremely low prevalence. We show the eff ectiveness of our one-shot approach for detecting face manipulations and compare it with other few-shot domain adaptation methods qualitatively and quantitative ly.
****

A Unified Optimization Framework for Low-Rank Inducing Penalties
Marcus Valtonen Ornhag, Carl Olsson; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8474-8483
In this paper we study the convex envelopes of a new class of functions. Using t his approach, we are able to unify two important classes of regularizers from un biased non-convex formulations and weighted nuclear norm penalties. This opens u p for possibilities of combining the best of both worlds, and to leverage each m ethods contribution to cases where simply enforcing one of the regularizers are insufficient. We show that the proposed regularizers can be incorporated in stan dard splitting schemes such as Alternating Direction Methods of Multipliers (ADM M), and other sub-gradient methods. This can be implemented efficiently since th

e the proximal operator can be computed fast. Furthermore, we show on real non-r igid structure from motion datasets, the issues that arise from using weighted n uclear norm penalties, and how this can be remedied using our proposed prior-fre e method.
****

Cost Volume Pyramid Based Depth Inference for Multi-View Stereo
Jiayu Yang, Wei Mao, Jose M. Alvarez, Miaomiao Liu; Proceedings of the IEEE/C VF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4877-4886
We propose a cost volume-based neural network for depth inference from multi-vie w images. We demonstrate that building a cost volume pyramid in a coarse-to-fine manner instead of constructing a cost volume at a fixed resolution leads to a c ompact, lightweight network and allows us inferring high resolution depth maps t o achieve better reconstruction results. To this end, we first build a cost volu me based on uniform sampling of fronto-parallel planes across the entire depth r ange at the coarsest resolution of an image. Then, given current depth estimate, we construct new cost volumes iteratively on the pixelwise depth residual to pe rform depth map refinement. While sharing similar insight with Point-MVSNet as p redicting and refining depth iteratively, we show that working on cost volume py ramid can lead to a more compact, yet efficient network structure compared with the Point-MVSNet on 3D points. We further provide detailed analyses of the relat ion between (residual) depth sampling and image resolution, which serves as a pr inciple for building compact cost volume pyramid. Experimental results on benchm ark datasets show that our model can perform 6x faster and has similar performan ce as state-of-the-art methods. Code is available at https://github.com/JiayuYAN G/CVP-MVSNet
****

Learning for Video Compression With Hierarchical Quality and Recurrent Enhanceme nt
Ren Yang, Fabian Mentzer, Luc Van Gool, Radu Timofte; Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 662 8-6637
In this paper, we propose a Hierarchical Learned Video Compression (HLVC) method with three hierarchical quality layers and a recurrent enhancement network. The frames in the first layer are compressed by an image compression method with th e highest quality. Using these frames as references, we propose the Bi-Direction al Deep Compression (BDDC) network to compress the second layer with relatively high quality. Then, the third layer frames are compressed with the lowest qualit y, by the proposed Single Motion Deep Compression (SMDC) network, which adopts a single motion map to estimate the motions of multiple frames, thus saving bits for motion information. In our deep decoder, we develop the Weighted Recurrent Q uality Enhancement (WRQE) network, which takes both compressed frames and the bi t stream as inputs. In the recurrent cell of WRQE, the memory and update signal are weighted by quality features to reasonably leverage multi-frame information for enhancement. In our HLVC approach, the hierarchical quality benefits the cod ing efficiency, since the high quality information facilitates the compression a nd enhancement of low quality frames at encoder and decoder sides, respectively. Finally, the experiments validate that our HLVC approach advances the state-of-the-art of deep video compression methods, and outperforms the "Low-Delay P (LDP) very fast" mode of x265 in terms of both PSNR and MS-SSIM. The project page is at https://github.com/RenYang-home/HLVC.
****

Sub-Frame Appearance and 6D Pose Estimation of Fast Moving Objects
Denys Rozumnyi, Jan Kotera, Filip Sroubek, Jiri Matas; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 67 78-6786
We propose a novel method that tracks fast moving objects, mainly non-uniform sp herical, in full 6 degrees of freedom, estimating simultaneously their 3D motion trajectory, 3D pose and object appearance changes with a time step that is a fr action of the video frame exposure time. The sub-frame object localization and a

ppearance estimation allows realistic temporal super-resolution and precise shap e estimation. The method, called TbD-3D (Tracking by Deblatting in 3D) relies on a novel reconstruction algorithm which solves a piece-wise deblurring and matti ng problem. The 3D rotation is estimated by minimizing the reprojection error. A s a second contribution, we present a new challenging dataset with fast moving o bjects that change their appearance and distance to the camera. High-speed camer a recordings with zero lag between frame exposures were used to generate videos with different frame rates annotated with ground-truth trajectory and pose.
****

TetraTSDF: 3D Human Reconstruction From a Single Image With a Tetrahedral Outer Shell

Hayato Onizuka, Zehra Hayirci, Diego Thomas, Akihiro Sugimoto, Hideaki Uchiy ama, Rin-ichiro Taniguchi; Proceedings of the IEEE/CVF Conference on Computer V ision and Pattern Recognition (CVPR), 2020, pp. 6011-6020

Recovering the 3D shape of a person from its 2D appearance is ill-posed due to a mbiguities. Nevertheless, with the help of convolutional neural networks (CNN) a nd prior knowledge on the 3D human body, it is possible to overcome such ambigui ties to recover detailed 3D shapes of human bodies from single images. Current s olutions, however, fail to reconstruct all the details of a person wearing loose clothes. This is because of either (a) huge memory requirement that cannot be m aintained even on modern GPUs or (b) the compact 3D representation that cannot e ncode all the details. In this paper, we propose the tetrahedral outer shell vol umetric truncated signed distance function (TetraTSDF) model for the human body, and its corresponding part connection network (PCN) for 3D human body shape reg ression. Our proposed model is compact, dense, accurate, and yet well suited for CNN-based regression task. Our proposed PCN allows us to learn the distribution of the TSDF in the tetrahedral volume from a single image in an end-to-end mann er. Results show that our proposed method allows to reconstruct detailed shapes of humans wearing loose clothes from single RGB images.
****

Flow2Stereo: Effective Self-Supervised Learning of Optical Flow and Stereo Match ing

Pengpeng Liu, Irwin King, Michael R. Lyu, Jia Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6648-66 57

In this paper, we propose a unified method to jointly learn optical flow and ste reo matching. Our first intuition is stereo matching can be modeled as a special case of optical flow, and we can leverage 3D geometry behind stereoscopic video s to guide the learning of these two forms of correspondences. We then enroll th is knowledge into the state-of-the-art self-supervised learning framework, and t rain one single network to estimate both flow and stereo. Second, we unveil the bottlenecks in prior self-supervised learning approaches, and propose to create a new set of challenging proxy tasks to boost performance. These two insights yi eld a single model that achieves the highest accuracy among all existing unsuper vised flow and stereo methods on KITTI 2012 and 2015 benchmarks. More remarkably , our self-supervised method even outperforms several state-of-the-art fully sup ervised methods, including PWC-Net and FlowNet2 on KITTI 2012.
****

Local Class-Specific and Global Image-Level Generative Adversarial Networks for Semantic-Guided Scene Generation

Hao Tang, Dan Xu, Yan Yan, Philip H.S. Torr, Nicu Sebe; Proceedings of the I EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7870-7879

In this paper, we address the task of semantic-guided scene generation. One open challenge widely observed in global image-level generation methods is the diffi culty of generating small objects and detailed local texture. To tackle this iss ue, in this work we consider learning the scene generation in a local context, a nd correspondingly design a local class-specific generative network with semanti c maps as a guidance, which separately constructs and learns sub-generators conc entrating on the generation of different classes, and is able to provide more sc

ene details. To learn more discriminative class-specific feature representations for the local generation, a novel classification module is also proposed. To combine the advantage of both global image-level and the local class-specific generation, a joint generation network is designed with an attention fusion module and a dual-discriminator structure embedded. Extensive experiments on two scene image generation tasks show superior generation performance of the proposed model. State-of-the-art results are established by large margins on both tasks and on challenging public benchmarks. The source code and trained models are available at https://github.com/Ha0Tang/LGGAN.
****

Fast Soft Color Segmentation
Naofumi Akimoto, Huachun Zhu, Yanghua Jin, Yoshimitsu Aoki; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8277-8286
We address the problem of soft color segmentation, defined as decomposing a given image into several RGBA layers, each containing only homogeneous color regions. The resulting layers from decomposition pave the way for applications that benefit from layer-based editing, such as recoloring and compositing of images and videos. The current state-of-the-art approach for this problem is hindered by slow processing time due to its iterative nature, and consequently does not scale to certain real-world scenarios. To address this issue, we propose a neural network based method for this task that decomposes a given image into multiple layers in a single forward pass. Furthermore, our method separately decomposes the color layers and the alpha channel layers. By leveraging a novel training objective, our method achieves proper assignment of colors amongst layers. As a consequence, our method achieve promising quality without existing issue of inference speed for iterative approaches. Our thorough experimental analysis shows that our method produces qualitative and quantitative results comparable to previous methods while achieving a 300,000x speed improvement. Finally, we utilize our proposed method on several applications, and demonstrate its speed advantage, especially in video editing.
****

Partial Weight Adaptation for Robust DNN Inference
Xiufeng Xie, Kyu-Han Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9573-9581
Mainstream video analytics uses a pre-trained DNN model with an assumption that inference input and training data follow the same probability distribution. However, this assumption does not always hold in the wild: autonomous vehicles may capture video with varying brightness; unstable wireless bandwidth calls for adaptive bitrate streaming of video; and, inference servers may serve inputs from heterogeneous IoT devices/cameras. In such situations, the level of input distortion changes rapidly, thus reshaping the probability distribution of the input. We present GearNN, an adaptive inference architecture that accommodates DNN inputs with varying distortions. GearNN employs an optimization algorithm to identify a tiny set of "distortion-sensitive" DNN parameters, given a memory budget. Based on the distortion level of the input, GearNN then adapts only the distortion-sensitive parameters, while reusing the rest of constant parameters across all input qualities. In our evaluation of DNN inference with dynamic input distortions, GearNN improves the accuracy (mIoU) by an average of 18.12% over a DNN trained with the undistorted dataset and 4.84% over stability training from Google, with only 1.8% extra memory overhead.
****

Deep Facial Non-Rigid Multi-View Stereo
Ziqian Bai, Zhaopeng Cui, Jamal Ahmed Rahim, Xiaoming Liu, Ping Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5850-5860
We present a method for 3D face reconstruction from multi-view images with different expressions. We formulate this problem from the perspective of non-rigid multi-view stereo (NRMVS). Unlike previous learning-based methods, which often regress the face shape directly, our method optimizes the 3D face shape by explicit

ly enforcing multi-view appearance consistency, which is known to be effective i
n recovering shape details according to conventional multi-view stereo methods.
Furthermore, by estimating face shape through optimization based on multi-view c
onsistency, our method can potentially have better generalization to unseen data
. However, this optimization is challenging since each input image has a differe
nt expression. We facilitate it with a CNN network that learns to regularize the
 non-rigid 3D face according to the input image and preliminary optimization res
ults. Extensive experiments show that our method achieves the state-of-the-art p
erformance on various datasets and generalizes well to in-the-wild data.
****

Deep Shutter Unrolling Network
Peidong Liu, Zhaopeng Cui, Viktor Larsson, Marc Pollefeys; Proceedings of the
 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp
. 5941-5949
We present a novel network for rolling shutter effect correction. Our network ta
kes two consecutive rolling shutter images and estimates the corresponding globa
l shutter image of the latest frame. The dense displacement field from a rolling
 shutter image to its corresponding global shutter image is estimated via a moti
on estimation network. The learned feature representation of a rolling shutter i
mage is then warped, via the displacement field, to its global shutter represent
ation by a differentiable forward warping block. An image decoder recovers the g
lobal shutter image based on the warped feature representation. Our network can
be trained end-to-end and only requires the global shutter image for supervision
. Since there is no public dataset available, we also propose two large datasets
: the Carla-RS dataset and the Fastec-RS dataset. Experimental results demonstra
te that our network outperforms the state-of-the-art methods. We make both our c
ode and datasets available at https://github.com/ethliup/DeepUnrollNet.
****

BlendMask: Top-Down Meets Bottom-Up for Instance Segmentation
Hao Chen, Kunyang Sun, Zhi Tian, Chunhua Shen, Yongming Huang, Youliang Yan
; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit
ion (CVPR), 2020, pp. 8573-8581
Instance segmentation is one of the fundamental vision tasks. Recently, fully co
nvolutional instance segmentation methods have drawn much attention as they are
often simpler and more efficient than two-stage approaches like Mask R-CNN. To d
ate, almost all such approaches fall behind the two-stage Mask R-CNN method in m
ask precision when models have similar computation complexity, leaving great roo
m for improvement. In this work, we achieve improved mask prediction by effectiv
ely combining instance-level information with semantic information with lower-le
vel fine-granularity. Our main contribution is a blender module which draws insp
iration from both top-down and bottom-up instance segmentation approaches. The p
roposed BlendMask can effectively predict dense per-pixel position-sensitive ins
tance features with very few channels, and learn attention maps for each instanc
e with merely one convolution layer, thus being fast in inference. BlendMask can
 be easily incorporate with the state-of-the-art one-stage detection frameworks
and outperforms Mask R-CNN under the same training schedule while being faster.
A light-weight version of BlendMask achieves 36.0 mAP at 27 FPS evaluated on a s
ingle 1080Ti. Because of its simplicity and efficacy, we hope that our BlendMask
 could serve as a simple yet strong baseline for a wide range of instance-wise p
rediction tasks.
****

Towards Learning Structure via Consensus for Face Segmentation and Parsing
Iacopo Masi, Joe Mathai, Wael AbdAlmageed; Proceedings of the IEEE/CVF Confere
nce on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5508-5518
Face segmentation is the task of densely labeling pixels on the face according t
o their semantics. While current methods place an emphasis on developing sophist
icated architectures, use conditional random fields for smoothness, or rather em
ploy adversarial training, we follow an alternative path towards robust face seg
mentation and parsing. Occlusions, along with other parts of the face, have a pr
oper structure that needs to be propagated in the model during training. Unlike

state-of-the-art methods that treat face segmentation as an independent pixel prediction problem, we argue instead that it should hold highly correlated outputs within the same object pixels. We thereby offer a novel learning mechanism to enforce structure in the prediction via consensus, guided by a robust loss function that forces pixel objects to be consistent with each other. Our face parser is trained by transferring knowledge from another model, yet it encourages spatial consistency while fitting the labels. Different than current practice, our method enjoys pixel-wise predictions, yet paves the way for fewer artifacts, less sparse masks, and spatially coherent outputs.
****

Pixel Consensus Voting for Panoptic Segmentation
Haochen Wang, Ruotian Luo, Michael Maire, Greg Shakhnarovich; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9464-9473
The core of our approach, Pixel Consensus Voting, is a framework for instance segmentation based on the generalized Hough transform. Pixels cast discretized, probabilistic votes for the likely regions that contain instance centroids. At the detected peaks that emerge in the voting heatmap, backprojection is applied to collect pixels and produce instance masks. Unlike a sliding window detector that densely enumerates object proposals, our method detects instances as a result of the consensus among pixel-wise votes. We implement vote aggregation and backprojection using native operators of a convolutional neural network. The discretization of centroid voting reduces the training of instance segmentation to pixel labeling, analogous and complementary to FCN-style semantic segmentation, leading to an efficient and unified architecture that jointly models things and stuff. We demonstrate the effectiveness of our pipeline on COCO and Cityscapes Panoptic Segmentation and obtain competitive results. Code will be open-sourced.
****

Towards Unsupervised Learning of Generative Models for 3D Controllable Image Synthesis
Yiyi Liao, Katja Schwarz, Lars Mescheder, Andreas Geiger; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5871-5880
In recent years, Generative Adversarial Networks have achieved impressive results in photorealistic image synthesis. This progress nurtures hopes that one day the classical rendering pipeline can be replaced by efficient models that are learned directly from images. However, current image synthesis models operate in the 2D domain where disentangling 3D properties such as camera viewpoint or object pose is challenging. Furthermore, they lack an interpretable and controllable representation. Our key hypothesis is that the image generation process should be modeled in 3D space as the physical world surrounding us is intrinsically three-dimensional. We define the new task of 3D controllable image synthesis and propose an approach for solving it by reasoning both in 3D space and in the 2D image domain. We demonstrate that our model is able to disentangle latent 3D factors of simple multi-object scenes in an unsupervised fashion from raw images. Compared to pure 2D baselines, it allows for synthesizing scenes that are consistent wrt. changes in viewpoint or object pose. We further evaluate various 3D representations in terms of their usefulness for this challenging task.
****

Exploit Clues From Views: Self-Supervised and Regularized Learning for Multiview Object Recognition
Chih-Hui Ho, Bo Liu, Tz-Ying Wu, Nuno Vasconcelos; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9090-9100
Multiview recognition has been well studied in the literature and achieves decent performance in object recognition and retrieval task. However, most previous works rely on supervised learning and some impractical underlying assumptions, such as the availability of all views in training and inference time. In this work, the problem of multiview self-supervised learning (MV-SSL) is investigated, where only image to object association is given. Given this setup, a novel surroga

te task for self-supervised learning is proposed by pursuing "object invariant" representation. This is solved by randomly selecting an image feature of an obje ct as object prototype, accompanied with multiview consistency regularization, w hich results in view invariant stochastic prototype embedding (VISPE). Experimen ts shows that the recognition and retrieval results using VISPE outperform that of other self-supervised learning methods on seen and unseen data. VISPE can als o be applied to semi-supervised scenario and demonstrates robust performance wit h limited data available. Code is available at https://github.com/chihhuiho/VISP E

****

SampleNet: Differentiable Point Cloud Sampling

Itai Lang, Asaf Manor, Shai Avidan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7578-7588

There is a growing number of tasks that work directly on point clouds. As the si ze of the point cloud grows, so do the computational demands of these tasks. A p ossible solution is to sample the point cloud first. Classic sampling approaches , such as farthest point sampling (FPS), do not consider the downstream task. A recent work showed that learning a task-specific sampling can improve results si gnificantly. However, the proposed technique did not deal with the non-different iability of the sampling operation and offered a workaround instead. We introduc e a novel differentiable relaxation for point cloud sampling that approximates s ampled points as a mixture of points in the primary input cloud. Our approximati on scheme leads to consistently good results on classification and geometry reco nstruction applications. We also show that the proposed sampling method can be u sed as a front to a point cloud registration network. This is a challenging task since sampling must be consistent across two different point clouds for a share d downstream task. In all cases, our approach outperforms existing non-learned a nd learned sampling alternatives. Our code is publicly available.

****

Guided Variational Autoencoder for Disentanglement Learning

Zheng Ding, Yifan Xu, Weijian Xu, Gaurav Parmar, Yang Yang, Max Welling, Z huowen Tu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7920-7929

We propose an algorithm, guided variational autoencoder (Guided-VAE), that is ab le to learn a controllable generative model by performing latent representation disentanglement learning. The learning objective is achieved by providing signal to the latent encoding/embedding in VAE without changing its main backbone arch itecture, hence retaining the desirable properties of the VAE. We design an unsu pervised and a supervised strategy in Guided-VAE and observe enhanced modeling a nd controlling capability over the vanilla VAE. In the unsupervised strategy, we guide the VAE learning by introducing a lightweight decoder that learns latent geometric transformation and principal components; in the supervised strategy, w e use an adversarial excitation and inhibition mechanism to encourage the disent anglement of the latent variables. Guided-VAE enjoys its transparency and simpli city for the general representation learning task, as well as disentanglement le arning. On a number of experiments for representation learning, improved synthes is/sampling, better disentanglement for classification, and reduced classificati on errors in meta learning have been observed.

****

Online Deep Clustering for Unsupervised Representation Learning

Xiaohang Zhan, Jiahao Xie, Ziwei Liu, Yew-Soon Ong, Chen Change Loy; Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6688-6697

Joint clustering and feature learning methods have shown remarkable performance in unsupervised representation learning. However, the training schedule alternat ing between feature clustering and network parameters update leads to unstable l earning of visual representations. To overcome this challenge, we propose Online Deep Clustering (ODC) that performs clustering and network update simultaneousl y rather than alternatingly. Our key insight is that the cluster centroids shoul d evolve steadily in keeping the classifier stably updated. Specifically, we des

ign and maintain two dynamic memory modules, i.e., samples memory to store sampl es' labels and features, and centroids memory for centroids evolution. We break down the abrupt global clustering into steady memory update and batch-wise label re-assignment. The process is integrated into network update iterations. In thi s way, labels and the network evolve shoulder-to-shoulder rather than alternatin gly. Extensive experiments demonstrate that ODC stabilizes the training process and boosts the performance effectively.
****

A Disentangling Invertible Interpretation Network for Explaining Latent Represen tations
Patrick Esser, Robin Rombach, Bjorn Ommer; Proceedings of the IEEE/CVF Confere nce on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9223-9232
Neural networks have greatly boosted performance in computer vision by learning powerful representations of input data. The drawback of end-to-end training for maximal overall performance are black-box models whose hidden representations ar e lacking interpretability: Since distributed coding is optimal for latent layer s to improve their robustness, attributing meaning to parts of a hidden feature vector or to individual neurons is hindered. We formulate interpretation as a tr anslation of hidden representations onto semantic concepts that are comprehensib le to the user. The mapping between both domains has to be bijective so that sem antic modifications in the target domain correctly alter the original representa tion. The proposed invertible interpretation network can be transparently applie d on top of existing architectures with no need to modify or retrain them. Conse quently, we translate an original representation to an equivalent yet interpreta ble one and backwards without affecting the expressiveness and performance of th e original. The invertible interpretation network disentangles the hidden repres entation into separate, semantically meaningful concepts. Moreover, we present a n efficient approach to define semantic concepts by only sketching two images an d also an unsupervised strategy. Experimental evaluation demonstrates the wide a pplicability to interpretation of existing classification and image generation n etworks as well as to semantically guided image manipulation.
****

SynSin: End-to-End View Synthesis From a Single Image
Olivia Wiles, Georgia Gkioxari, Richard Szeliski, Justin Johnson; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2 020, pp. 7467-7477
View synthesis allows for the generation of new views of a scene given one or mo re images. This is challenging; it requires comprehensively understanding the 3D scene from images. As a result, current methods typically use multiple images, train on ground-truth depth, or are limited to synthetic data. We propose a nove l end-to-end model for this task using a single image at test time; it is traine d on real images without any ground-truth 3D information. To this end, we introd uce a novel differentiable point cloud renderer that is used to transform a late nt 3D point cloud of features into the target view. The projected features are d ecoded by our refinement network to inpaint missing regions and generate a reali stic output image. The 3D component inside of our generative model allows for in terpretable manipulation of the latent feature space at test time, e.g. we can a nimate trajectories from a single image. Additionally, we can generate high reso lution images and generalise to other input resolutions. We outperform baselines and prior work on the Matterport, Replica, and RealEstate10K datasets.
****

HOPE-Net: A Graph-Based Model for Hand-Object Pose Estimation
Bardia Doosti, Shujon Naha, Majid Mirbagheri, David J. Crandall; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20 20, pp. 6608-6617
Hand-object pose estimation (HOPE) aims to jointly detect the poses of both a ha nd and of a held object. In this paper, we propose a lightweight model called HO PE-Net which jointly estimates hand and object pose in 2D and 3D in real-time. O ur network uses a cascade of two adaptive graph convolutional neural networks, o ne to estimate 2D coordinates of the hand joints and object corners, followed by

another to convert 2D coordinates to 3D. Our experiments show that through end-to-end training of the full network, we achieve better accuracy for both the 2D and 3D coordinate estimation problems. The proposed 2D to 3D graph convolution-based model could be applied to other 3D landmark detection problems, where it is possible to first predict the 2D keypoints and then transform them to 3D.
****

Auto-Tuning Structured Light by Optical Stochastic Gradient Descent
Wenzheng Chen, Parsa Mirdehghan, Sanja Fidler, Kiriakos N. Kutulakos; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5970-5980
We consider the problem of optimizing the performance of an active imaging system by automatically discovering the illuminations it should use, and the way to decode them. Our approach tackles two seemingly incompatible goals: (1) "tuning" the illuminations and decoding algorithm precisely to the devices at hand---to their optical transfer functions, non-linearities, spectral responses, image processing pipelines---and (2) doing so without modeling or calibrating the system; without modeling the scenes of interest; and without prior training data. The key idea is to formulate a stochastic gradient descent (SGD) optimization procedure that puts the actual system in the loop: projecting patterns, capturing images, and calculating the gradient of expected reconstruction error. We apply this idea to structured-light triangulation to "auto-tune" several devices---from smartphones and laser projectors to advanced computational cameras. Our experiments show that despite being model-free and automatic, optical SGD can boost system 3D accuracy substantially over state-of-the-art coding schemes.
****

HandVoxNet: Deep Voxel-Based Network for 3D Hand Shape and Pose Estimation From a Single Depth Map
Jameel Malik, Ibrahim Abdelaziz, Ahmed Elhayek, Soshi Shimada, Sk Aziz Ali, Vladislav Golyanik, Christian Theobalt, Didier Stricker; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7113-7122
3D hand shape and pose estimation from a single depth map is a new and challenging computer vision problem with many applications. The state-of-the-art methods directly regress 3D hand meshes from 2D depth images via 2D convolutional neural networks, which leads to artefacts in the estimations due to perspective distortions in the images. In contrast, we propose a novel architecture with 3D convolutions trained in a weakly-supervised manner. The input to our method is a 3D voxelized depth map, and we rely on two hand shape representations. The first one is the 3D voxelized grid of the shape which is accurate but does not preserve the mesh topology and the number of mesh vertices. The second representation is the 3D hand surface which is less accurate but does not suffer from the limitations of the first representation. We combine the advantages of these two representations by registering the hand surface to the voxelized hand shape. In the extensive experiments, the proposed approach improves over the state of the art by47.8 % on the SynHand5M dataset. Moreover, our augmentation policy for voxelized depth maps further enhances the accuracy of 3D hand pose estimation on real data. Our method produces visually more reasonable and realistic hand shapes on NYU and BigHand2.2M datasets compared to the existing approaches.
****

Deep 3D Portrait From a Single Image
Sicheng Xu, Jiaolong Yang, Dong Chen, Fang Wen, Yu Deng, Yunde Jia, Xin Tong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7710-7720
In this paper, we present a learning-based approach for recovering the 3D geometry of human head from a single portrait image. Our method is learned in an unsupervised manner without any ground-truth 3D data. We represent the head geometry with a parametric 3D face model together with a depth map for other head regions including hair and ear. A two-step geometry learning scheme is proposed to learn 3D head reconstruction from in-the-wild face images, where we first learn face shape on single images using self-reconstruction and then learn hair and ear ge

ometry using pairs of images in a stereo-matching fashion. The second step is based on the output of the first to not only improve the accuracy but also ensure the consistency of overall head geometry. We evaluate the accuracy of our method both in 3D and with pose manipulation tasks on 2D images. We alter pose based on the recovered geometry and apply a refinement network trained with adversarial learning to ameliorate the reprojected images and translate them to the real image domain. Extensive evaluations and comparison with previous methods show that our new method can produce high-fidelity 3D head geometry and head pose manipulation results.
****

AnimalWeb: A Large-Scale Hierarchical Dataset of Annotated Animal Faces
Muhammad Haris Khan, John McDonagh, Salman Khan, Muhammad Shahabuddin, Aditya Arora, Fahad Shahbaz Khan, Ling Shao, Georgios Tzimiropoulos; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6939-6948
Several studies show that animal needs are often expressed through their faces. Though remarkable progress has been made towards the automatic understanding of human faces, this has not been the case with animal faces. There exists significant room for algorithmic advances that could realize automatic systems for interpreting animal faces. Besides scientific value, resulting technology will foster better and cheaper animal care. We believe the underlying research progress is mainly obstructed by the lack of an adequately annotated dataset of animal faces, covering a wide spectrum of animal species. To this end, we introduce a large-scale, hierarchical annotated dataset of animal faces, featuring 22.4K faces from 350 diverse species and 21 animal orders across biological taxonomy. These faces are captured `in-the-wild' conditions and are consistently annotated with 9 landmarks on key facial features. The dataset is structured and scalable by design; its development underwent four systematic stages involving rigorous, overall effort of over 6K man-hours. We benchmark it for face alignment using the existing art under two new problem settings. Results showcase its challenging nature, unique attributes and present definite prospects for novel, adaptive, and generalized face-oriented CV algorithms. Further benchmarking the dataset across face detection and fine-grained recognition tasks demonstrates its multi-task applications and room for improvement. The dataset is available at: https://fdmaproject.wordpress.com/.
****

MANTRA: Memory Augmented Networks for Multiple Trajectory Prediction
Francesco Marchetti, Federico Becattini, Lorenzo Seidenari, Alberto Del Bimbo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7143-7152
Autonomous vehicles are expected to drive in complex scenarios with several independent non cooperating agents. Path planning for safely navigating in such environments can not just rely on perceiving present location and motion of other agents. It requires instead to predict such variables in a far enough future. In this paper we address the problem of multimodal trajectory prediction exploiting a Memory Augmented Neural Network. Our method learns past and future trajectory embeddings using recurrent neural networks and exploits an associative external memory to store and retrieve such embeddings. Trajectory prediction is then performed by decoding in-memory future encodings conditioned with the observed past. We incorporate scene knowledge in the decoding state by learning a CNN on top of semantic scene maps. Memory growth is limited by learning a writing controller based on the predictive capability of existing embeddings. We show that our method is able to natively perform multi-modal trajectory prediction obtaining state-of-the art results on three datasets. Moreover, thanks to the non-parametric nature of the memory module, we show how once trained our system can continuously improve by ingesting novel patterns.
****

Neural Point Cloud Rendering via Multi-Plane Projection
Peng Dai, Yinda Zhang, Zhuwen Li, Shuaicheng Liu, Bing Zeng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020,

pp. 7830-7839
We present a new deep point cloud rendering pipeline through multi-plane projections. The input to the network is the raw point cloud of a scene and the output are image or image sequences from a novel view or along a novel camera trajectory. Unlike previous approaches that directly project features from 3D points onto 2D image domain, we propose to project these features into a layered volume of camera frustum. In this way, the visibility of 3D points can be automatically learnt by the network, such that ghosting effects due to false visibility check as well as occlusions caused by noise interferences are both avoided successfully. Next, the 3D feature volume is fed into a 3D CNN to produce multiple planes of images w.r.t. the space division in the depth directions. The multi-plane images are then blended based on learned weights to produce the final rendering results. Experiments show that our network produces more stable renderings compared to previous methods, especially near the object boundaries. Moreover, our pipeline is robust to noisy and relatively sparse point cloud for a variety of challenging scenes.
****

A2dele: Adaptive and Attentive Depth Distiller for Efficient RGB-D Salient Object Detection
Yongri Piao, Zhengkun Rong, Miao Zhang, Weisong Ren, Huchuan Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9060-9069
Existing state-of-the-art RGB-D salient object detection methods explore RGB-D data relying on a two-stream architecture, in which an independent subnetwork is required to process depth data. This inevitably incurs extra computational costs and memory consumption, and using depth data during testing may hinder the practical applications of RGB-D saliency detection. To tackle these two dilemmas, we propose a depth distiller (A2dele) to explore the way of using network prediction and attention as two bridges to transfer the depth knowledge from the depth stream to the RGB stream. First, by adaptively minimizing the differences between predictions generated from the depth stream and RGB stream, we realize the desired control of pixel-wise depth knowledge transferred to the RGB stream. Second, to transfer the localization knowledge to RGB features, we encourage consistencies between the dilated prediction of the depth stream and the attention map from the RGB stream. As a result, we achieve a lightweight architecture without use of depth data at test time by embedding our A2dele. Our extensive experimental evaluation on five benchmarks demonstrate that our RGB stream achieves state-of-the-art performance, which tremendously minimizes the model size by 76% and runs 12 times faster, compared with the best performing method. Furthermore, our A2dele can be applied to existing RGB-D networks to significantly improve their efficiency while maintaining performance (boosts FPS by nearly twice for DMRA and 3 times for CPFP).
****

Continual Learning With Extended Kronecker-Factored Approximate Curvature
Janghyeon Lee, Hyeong Gwon Hong, Donggyu Joo, Junmo Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9001-9010
We propose a quadratic penalty method for continual learning of neural networks that contain batch normalization (BN) layers. The Hessian of a loss function represents the curvature of the quadratic penalty function, and a Kronecker-factored approximate curvature (K-FAC) is used widely to practically compute the Hessian of a neural network. However, the approximation is not valid if there is dependence between examples, typically caused by BN layers in deep network architectures. We extend the K-FAC method so that the inter-example relations are taken into account and the Hessian of deep neural networks can be properly approximated under practical assumptions. We also propose a method of weight merging and reparameterization to properly handle statistical parameters of BN, which plays a critical role for continual learning with BN, and a method that selects hyperparameters without source task data. Our method shows better performance than baselines in the permuted MNIST task with BN layers and in sequential learning from the

ImageNet classification task to fine-grained classification tasks with ResNet-50, without any explicit or implicit use of source task data for hyperparameter selection.

****

Domain Balancing: Face Recognition on Long-Tailed Domains

Dong Cao, Xiangyu Zhu, Xingyu Huang, Jianzhu Guo, Zhen Lei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5671-5679

Long-tailed problem has been an important topic in face recognition task. However, existing methods only concentrate on the long-tailed distribution of classes. Differently, we devote to the long-tailed domain distribution problem, which refers to the fact that a small number of domains frequently appear while other domains far less existing. The key challenge of the problem is that domain labels are too complicated (related to race, age, pose, illumination, etc.) and inaccessible in real applications. In this paper, we propose a novel Domain Balancing (DB) mechanism to handle this problem. Specifically, we first propose a Domain Frequency Indicator (DFI) to judge whether a sample is from head domains or tail domains. Secondly, we formulate a light-weighted Residual Balancing Mapping (RBM) block to balance the domain distribution by adjusting the network according to DFI. Finally, we propose a Domain Balancing Margin (DBM) in the loss function to further optimize the feature space of the tail domains to improve generalization. Extensive analysis and experiments on several face recognition benchmarks demonstrate that the proposed method effectively enhances the generalization capacities and achieves superior performance.

****

Neural Pose Transfer by Spatially Adaptive Instance Normalization

Jiashun Wang, Chao Wen, Yanwei Fu, Haitao Lin, Tianyun Zou, Xiangyang Xue, Yinda Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5831-5839

Pose transfer has been studied for decades, in which the pose of a source mesh is applied to a target mesh. Particularly in this paper, we are interested in transferring the pose of source human mesh to deform the target human mesh, while the source and target meshes may have different identity information. Traditional studies assume that the paired source and target meshes are existed with the point-wise correspondences of user annotated landmarks/mesh points, which requires heavy labelling efforts. On the other hand, the generalization ability of deep models is limited, when the source and target meshes have different identities. To break this limitation, we proposes the first neural pose transfer model that solves the pose transfer via the latest technique for image style transfer, leveraging the newly proposed component -- spatially adaptive instance normalization. Our model does not require any correspondences between the source and target meshes. Extensive experiments show that the proposed model can effectively transfer deformation from source to target meshes, and has good generalization ability to deal with unseen identities or poses of meshes. Code is available at https://github.com/jiashunwang/Neural-Pose-Transfer.

****

RoutedFusion: Learning Real-Time Depth Map Fusion

Silvan Weder, Johannes Schonberger, Marc Pollefeys, Martin R. Oswald; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4887-4897

The efficient fusion of depth maps is a key part of most state-of-the-art 3D reconstruction methods. Besides requiring high accuracy, these depth fusion methods need to be scalable and real-time capable. To this end, we present a novel real-time capable machine learning-based method for depth map fusion. Similar to the seminal depth map fusion approach by Curless and Levoy, we only update a local group of voxels to ensure real-time capability. Instead of a simple linear fusion of depth information, we propose a neural network that predicts non-linear updates to better account for typical fusion errors. Our network is composed of a 2D depth routing network and a 3D depth fusion network which efficiently handle sensor-specific noise and outliers. This is especially useful for surface edges a

nd thin objects for which the original approach suffers from thickening artifacts. Our method outperforms the traditional fusion approach and related learned approaches on both synthetic and real data. We demonstrate the performance of our method in reconstructing fine geometric details from noise and outlier contaminated data on various scenes.
****

Coherent Reconstruction of Multiple Humans From a Single Image
Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, Kostas Daniilidis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5579-5588
In this work, we address the problem of multi-person 3D pose estimation from a single image. A typical regression approach in the top-down setting of this problem would first detect all humans and then reconstruct each one of them independently. However, this type of prediction suffers from incoherent results, e.g., interpenetration and inconsistent depth ordering between the people in the scene. Our goal is to train a single network that learns to avoid these problems and generate a coherent 3D reconstruction of all the humans in the scene. To this end, a key design choice is the incorporation of the SMPL parametric body model in our top-down framework, which enables the use of two novel losses. First, a distance field-based collision loss penalizes interpenetration among the reconstructed people. Second, a depth ordering-aware loss reasons about occlusions and promotes a depth ordering of people that leads to a rendering which is consistent with the annotated instance segmentation. This provides depth supervision signals to the network, even if the image has no explicit 3D annotations. The experiments show that our approach outperforms previous methods on standard 3D pose benchmarks, while our proposed losses enable more coherent reconstruction in natural images. The project website with videos, results, and code can be found at: https://jiangwenpl.github.io/multiperson
****

High-Performance Long-Term Tracking With Meta-Updater
Kenan Dai, Yunhua Zhang, Dong Wang, Jianhua Li, Huchuan Lu, Xiaoyun Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6298-6307
Long-term visual tracking has drawn increasing attention because it is much closer to practical applications than short-term tracking. Most top-ranked long-term trackers adopt the offline-trained Siamese architectures, thus,they cannot benefit from great progress of short-term trackers with online update. However, it is quite risky to straightforwardly introduce online-update-based trackers to solve the long-term problem, due to long-term uncertain and noisy observations. In this work, we propose a novel offline-trained Meta-Updater to address an important but unsolved problem: Is the tracker ready for updating in the current frame? The proposed meta-updater can effectively integrate geometric, discriminative, and appearance cues in a sequential manner, and then mine the sequential information with a designed cascaded LSTM module. Our meta-updater learns a binary output to guide the tracker's update and can be easily embedded into different trackers. This work also introduces a long-term tracking framework consisting of an online local tracker, an online verifier, a SiamRPN-based re-detector, and our meta-updater. Numerous experimental results on the VOT2018LT,VOT2019LT, OxUvALT, TLP, and LaSOT benchmarks show that our tracker performs remarkably better than other competing algorithms. Our project is available on the website: https://github.com/Daikenan/LTMU.
****

Rethinking Class-Balanced Methods for Long-Tailed Visual Recognition From a Domain Adaptation Perspective
Muhammad Abdullah Jamal, Matthew Brown, Ming-Hsuan Yang, Liqiang Wang, Boqing Gong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7610-7619
Object frequency in the real world often follows a power law, leading to a mismatch between datasets with long-tailed class distributions seen by a machine learning model and our expectation of the model to perform well on all classes. We a

nalyze this mismatch from a domain adaptation point of view. First of all, we co nnect existing class-balanced methods for long-tailed classification to target s hift, a well-studied scenario in domain adaptation. The connection reveals that these methods implicitly assume that the training data and test data share the s ame class-conditioned distribution, which does not hold in general and especiall y for the tail classes. While a head class could contain abundant and diverse tr aining examples that well represent the expected data at inference time, the tai l classes are often short of representative training data. To this end, we propo se to augment the classic class-balanced learning by explicitly estimating the d ifferences between the class-conditioned distributions with a meta-learning appr oach. We validate our approach with six benchmark datasets and three loss functi ons.
****

Softmax Splatting for Video Frame Interpolation
Simon Niklaus, Feng Liu; Proceedings of the IEEE/CVF Conference on Computer Vis ion and Pattern Recognition (CVPR), 2020, pp. 5437-5446
Differentiable image sampling in the form of backward warping has seen broad ado ption in tasks like depth estimation and optical flow prediction. In contrast, h ow to perform forward warping has seen less attention, partly due to additional challenges such as resolving the conflict of mapping multiple pixels to the same target location in a differentiable way. We propose softmax splatting to addres s this paradigm shift and show its effectiveness on the application of frame int erpolation. Specifically, given two input frames, we forward-warp the frames and their feature pyramid representations based on an optical flow estimate using s oftmax splatting. In doing so, the softmax splatting seamlessly handles cases wh ere multiple source pixels map to the same target location. We then use a synthe sis network to predict the interpolation result from the warped representations. Our softmax splatting allows us to not only interpolate frames at an arbitrary time but also to fine tune the feature pyramid and the optical flow. We show tha t our synthesis approach, empowered by softmax splatting, achieves new state-of- the-art results for video frame interpolation.
****

Cross-Domain Correspondence Learning for Exemplar-Based Image Translation
Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, Fang Wen; Proceedings of the IEEE/C VF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5143- 5153
We present a general framework for exemplar-based image translation, which synth esizes a photo-realistic image from the input in a distinct domain (e.g., semant ic segmentation mask, or edge map, or pose keypoints), given an exemplar image. The output has the style (e.g., color, texture) in consistency with the semantic ally corresponding objects in the exemplar. We propose to jointly learn the cros s-domain correspondence and the image translation, where both tasks facilitate e ach other and thus can be learned with weak supervision. The images from distinc t domains are first aligned to an intermediate domain where dense correspondence is established. Then, the network synthesizes images based on the appearance of semantically corresponding patches in the exemplar. We demonstrate the effectiv eness of our approach in several image translation tasks. Our method is superior to state-of-the-art methods in terms of image quality significantly, with the i mage style faithful to the exemplar with semantic consistency. Moreover, we show the utility of our method for several applications.
****

A Multi-Task Mean Teacher for Semi-Supervised Shadow Detection
Zhihao Chen, Lei Zhu, Liang Wan, Song Wang, Wei Feng, Pheng-Ann Heng; Proce edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV PR), 2020, pp. 5611-5620
Existing shadow detection methods suffer from an intrinsic limitation in relying on limited labeled datasets, and they may produce poor results in some complica ted situations. To boost the shadow detection performance, this paper presents a multi-task mean teacher model for semi-supervised shadow detection by leveragin g unlabeled data and exploring the learning of multiple information of shadows s

imultaneously. To be specific, we first build a multi-task baseline model to sim ultaneously detect shadow regions, shadow edges, and shadow count by leveraging their complementary information and assign this baseline model to the student an d teacher network. After that, we encourage the predictions of the three tasks f rom the student and teacher networks to be consistent for computing a consistenc y loss on unlabeled data, which is then added to the supervised loss on the labe led data from the predictions of the multi-task baseline model. Experimental res ults on three widely-used benchmark datasets show that our method consistently o utperforms all the compared state-of- the-art methods, which verifies that the p roposed network can effectively leverage additional unlabeled data to boost the shadow detection performance.
****

Closed-Loop Matters: Dual Regression Networks for Single Image Super-Resolution
Yong Guo, Jian Chen, Jingdong Wang, Qi Chen, Jiezhang Cao, Zeshuai Deng, Y anwu Xu, Mingkui Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5407-5416
Deep neural networks have exhibited promising performance in image super-resolut ion (SR) by learning a nonlinear mapping function from low-resolution (LR) image s to high-resolution (HR) images. However, there are two underlying limitations to existing SR methods. First, learning the mapping function from LR to HR image s is typically an ill-posed problem, because there exist infinite HR images that can be downsampled to the same LR image. As a result, the space of the possible functions can be extremely large, which makes it hard to find a good solution. Second, the paired LR-HR data may be unavailable in real-world applications and the underlying degradation method is often unknown. For such a more general case , existing SR models often incur the adaptation problem and yield poor performan ce. To address the above issues, we propose a dual regression scheme by introduc ing an additional constraint on LR data to reduce the space of the possible func tions. Specifically, besides the mapping from LR to HR images, we learn an addit ional dual regression mapping estimates the down-sampling kernel and reconstruct LR images, which forms a closed-loop to provide additional supervision. More cr itically, since the dual regression process does not depend on HR images, we can directly learn from LR images. In this sense, we can easily adapt SR models to real-world data, e.g., raw video frames from YouTube. Extensive experiments with paired training data and unpaired real-world data demonstrate our superiority o ver existing methods.
****

ROAM: Recurrently Optimizing Tracking Model
Tianyu Yang, Pengfei Xu, Runbo Hu, Hua Chai, Antoni B. Chan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6718-6727
In this paper, we design a tracking model consisting of response generation and bounding box regression, where the first component produces a heat map to indica te the presence of the object at different positions and the second part regress es the relative bounding box shifts to anchors mounted on sliding-window locatio ns. Thanks to the resizable convolutional filters used in both components to ada pt to the shape changes of objects, our tracking model does not need to enumerat e different sized anchors, thus saving model parameters. To effectively adapt th e model to appearance variations, we propose to offline train a recurrent neural optimizer to update tracking model in a meta-learning setting, which can conver ge the model in a few gradient steps. This improves the convergence speed of upd ating the tracking model while achieving better performance. We extensively eval uate our trackers, ROAM and ROAM++, on the OTB, VOT, LaSOT, GOT-10K and Tracking Net benchmark and our methods perform favorably against state-of-the-art algorit hms.
****

Wavelet Integrated CNNs for Noise-Robust Image Classification
Qiufu Li, Linlin Shen, Sheng Guo, Zhihui Lai; Proceedings of the IEEE/CVF Con ference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7245-7254
Convolutional Neural Networks (CNNs) are generally prone to noise interruptions,

i.e., small image noise can cause drastic changes in the output. To suppress th e noise effect to the final predication, we enhance CNNs by replacing max-poolin g, strided-convolution, and average-pooling with Discrete Wavelet Transform (DWT). We present general DWT and Inverse DWT (IDWT) layers applicable to various wa velets like Haar, Daubechies, and Cohen, etc., and design wavelet integrated CNN s (WaveCNets) using these layers for image classification. In WaveCNets, feature maps are decomposed into the low-frequency and high-frequency components during the down-sampling. The low-frequency component stores main information includin g the basic object structures, which is transmitted into the subsequent layers t o extract robust high-level features. The high-frequency components, containing most of the data noise, are dropped during inference to improve the noise-robust ness of the WaveCNets. Our experimental results on ImageNet and ImageNet-C (the noisy version of ImageNet) show that WaveCNets, the wavelet integrated versions of VGG, ResNets, and DenseNet, achieve higher accuracy and better noise-robustne ss than their vanilla versions.

****

Towards Causal VQA: Revealing and Reducing Spurious Correlations by Invariant an d Covariant Semantic Editing

Vedika Agarwal, Rakshith Shetty, Mario Fritz; Proceedings of the IEEE/CVF Conf erence on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9690-9698

Despite significant success in Visual Question Answering (VQA), VQA models have been shown to be notoriously brittle to linguistic variations in the questions. Due to deficiencies in models and datasets, today's models often rely on correla tions rather than predictions that are causal w.r.t. data. In this paper, we pro pose a novel way to analyze and measure the robustness of the state of the art m odels w.r.t semantic visual variations as well as propose ways to make models mo re robust against spurious correlations. Our method performs automated semantic image manipulations and tests for consistency in model predictions to quantify t he model robustness as well as generate synthetic data to counter these problems . We perform our analysis on three diverse, state of the art VQA models and dive rse question types with a particular focus on challenging counting questions. In addition, we show that models can be made significantly more robust against inc onsistent predictions using our edited data. Finally, we show that results also translate to real-world error cases of state of the art models, which results in improved overall performance

****

FReeNet: Multi-Identity Face Reenactment

Jiangning Zhang, Xianfang Zeng, Mengmeng Wang, Yusu Pan, Liang Liu, Yong Li u, Yu Ding, Changjie Fan; Proceedings of the IEEE/CVF Conference on Computer V ision and Pattern Recognition (CVPR), 2020, pp. 5326-5335

This paper presents a novel multi-identity face reenactment framework, named FRe eNet, to transfer facial expressions from an arbitrary source face to a target f ace with a shared model. The proposed FReeNet consists of two parts: Unified Lan dmark Converter (ULC) and Geometry-aware Generator (GAG). The ULC adopts an enco de-decoder architecture to efficiently convert expression in a latent landmark s pace, which significantly narrows the gap of the face contour between source and target identities. The GAG leverages the converted landmark to reenact the phot orealistic image with a reference image of the target person. Moreover, a new tr iplet perceptual loss is proposed to force the GAG module to learn appearance an d geometry information simultaneously, which also enriches facial details of the reenacted images. Further experiments demonstrate the superiority of our approa ch for generating photorealistic and expression-alike faces, as well as the flex ibility for transferring facial expressions between identities.

****

Deep Snake for Real-Time Instance Segmentation

Sida Peng, Wen Jiang, Huaijin Pi, Xiuli Li, Hujun Bao, Xiaowei Zhou; Procee dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP R), 2020, pp. 8533-8542

This paper introduces a novel contour-based approach named deep snake for real-t ime instance segmentation. Unlike some recent methods that directly regress the

coordinates of the object boundary points from an image, deep snake uses a neural network to iteratively deform an initial contour to match the object boundary, which implements the classic idea of snake algorithms with a learning-based approach. For structured feature learning on the contour, we propose to use circular convolution in deep snake, which better exploits the cycle-graph structure of a contour compared against generic graph convolution. Based on deep snake, we develop a two-stage pipeline for instance segmentation: initial contour proposal and contour deformation, which can handle errors in object localization. Experiments show that the proposed approach achieves competitive performances on the Cityscapes, KINS, SBD and COCO datasets while being efficient for real-time applications with a speed of 32.3 fps for 512 x 512 images on a 1080Ti GPU. The code is available at https://github.com/zju3dv/snake/.

****

Learning Identity-Invariant Motion Representations for Cross-ID Face Reenactment
Po-Hsiang Huang, Fu-En Yang, Yu-Chiang Frank Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7084-7092

Human face reenactment aims at transferring motion patterns from one face (from a source-domain video) to an-other (in the target domain with the identity of interest).While recent works report impressive results, they are notable to handle multiple identities in a unified model. In this paper, we propose a unique network of CrossID-GAN to perform multi-ID face reenactment. Given a source-domain video with extracted facial landmarks and a target-domain image, our CrossID-GAN learns the identity-invariant motion patterns via the extracted landmarks and such information to produce the videos whose ID matches that of the target domain. Both supervised and unsupervised settings are proposed to train and guide our model during training.Our qualitative/quantitative results confirm the robustness and effectiveness of our model, with ablation studies confirming our network design.

****

Unsupervised Domain Adaptation via Structurally Regularized Deep Clustering
Hui Tang, Ke Chen, Kui Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8725-8735

Unsupervised domain adaptation (UDA) is to make predictions for unlabeled data on a target domain, given labeled data on a source domain whose distribution shifts from the target one. Mainstream UDA methods learn aligned features between the two domains, such that a classifier trained on the source features can be readily applied to the target ones. However, such a transferring strategy has a potential risk of damaging the intrinsic discrimination of target data. To alleviate this risk, we are motivated by the assumption of structural domain similarity, and propose to directly uncover the intrinsic target discrimination via discriminative clustering of target data. We constrain the clustering solutions using structural source regularization that hinges on our assumed structural domain similarity. Technically, we use a flexible framework of deep network based discriminative clustering that minimizes the KL divergence between predictive label distribution of the network and an introduced auxiliary one; replacing the auxiliary distribution with that formed by ground-truth labels of source data implements the structural source regularization via a simple strategy of joint network training. We term our proposed method as Structurally Regularized Deep Clustering (SRDC), where we also enhance target discrimination with clustering of intermediate network features, and enhance structural regularization with soft selection of less divergent source examples. Careful ablation studies show the efficacy of our proposed SRDC. Notably, with no explicit domain alignment, SRDC outperforms all existing methods on three UDA benchmarks.

****

Augment Your Batch: Improving Generalization Through Instance Repetition
Elad Hoffer, Tal Ben-Nun, Itay Hubara, Niv Giladi, Torsten Hoefler, Daniel Soudry; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8129-8138

Large-batch SGD is important for scaling training of deep neural networks. Howev

er, without fine-tuning hyperparameter schedules, the generalization of the mode l may be hampered. We propose to use batch augmentation: replicating instances o f samples within the same batch with different data augmentations. Batch augment ation acts as a regularizer and an accelerator, increasing both generalization a nd performance scaling for a fixed budget of optimization steps. We analyze the effect of batch augmentation on gradient variance and show that it empirically i mproves convergence for a wide variety of networks and datasets. Our results sho w that batch augmentation reduces the number of necessary SGD updates to achieve the same accuracy as the state-of-the-art. Overall, this simple yet effective m ethod enables faster training and better generalization by allowing more computa tional resources to be used concurrently.

****

AdaCoF: Adaptive Collaboration of Flows for Video Frame Interpolation

Hyeongmin Lee, Taeoh Kim, Tae-young Chung, Daehyun Pak, Yuseok Ban, Sangyou n Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec ognition (CVPR), 2020, pp. 5316-5325

Video frame interpolation is one of the most challenging tasks in video processi ng research. Recently, many studies based on deep learning have been suggested. Most of these methods focus on finding locations with useful information to esti mate each output pixel using their own frame warping operations. However, many o f them have Degrees of Freedom (DoF) limitations and fail to deal with the compl ex motions found in real world videos. To solve this problem, we propose a new w arping module named Adaptive Collaboration of Flows (AdaCoF). Our method estimat es both kernel weights and offset vectors for each target pixel to synthesize th e output frame. AdaCoF is one of the most generalized warping modules compared t o other approaches, and covers most of them as special cases of it. Therefore, i t can deal with a significantly wide domain of complex motions. To further impro ve our framework and synthesize more realistic outputs, we introduce dual-frame adversarial loss which is applicable only to video frame interpolation tasks. Th e experimental results show that our method outperforms the state-of-the-art met hods for both fixed training set environments and the Middlebury benchmark. Our source code is available at https://github.com/HyeongminLEE/AdaCoF-pytorch

****

Blurry Video Frame Interpolation

Wang Shen, Wenbo Bao, Guangtao Zhai, Li Chen, Xiongkuo Min, Zhiyong Gao; Pr oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5114-5123

Existing works reduce motion blur and up-convert frame rate through two separate ways, including frame deblurring and frame interpolation. However, few studies have approached the joint video enhancement problem, namely synthesizing high-fr ame-rate clear results from low-frame-rate blurry inputs. In this paper, we prop ose a blurry video frame interpolation method to reduce motion blur and up-conve rt frame rate simultaneously. Specifically, we develop a pyramid module to cycli cally synthesize clear intermediate frames. The pyramid module features adjustab le spatial receptive field and temporal scope, thus contributing to controllable computational complexity and restoration ability. Besides, we propose an inter-pyramid recurrent module to connect sequential models to exploit the temporal re lationship. The pyramid module integrates a recurrent module, thus can iterative ly synthesize temporally smooth results without significantly increasing the mod el size. Extensive experimental results demonstrate that our method performs fav orably against state-of-the-art methods. The source code and pre-trained model a re available at https://github.com/laomao0/BIN.

****

Self-Learning With Rectification Strategy for Human Parsing

Tao Li, Zhiyuan Liang, Sanyuan Zhao, Jiahao Gong, Jianbing Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2 020, pp. 9263-9272

In this paper, we solve the sample shortage problem in the human parsing task. W e begin with the self-learning strategy, which generates pseudo-labels for unlab eled data to retrain the model. However, directly using noisy pseudo-labels will

cause error amplification and accumulation. Considering the topology structure of human body, we propose a trainable graph reasoning method that establishes internal structural connections between graph nodes to correct two typical errors in the pseudo-labels, i.e., the global structural error and the local consistency error. For the global error, we first transform category-wise features into a high-level graph model with coarse-grained structural information, and then decouple the high-level graph to reconstruct the category features. The reconstructed features have a stronger ability to represent the topology structure of the human body. Enlarging the receptive field of features can effectively reducing the local error. We first project feature pixels into a local graph model to capture pixel-wise relations in a hierarchical graph manner, then reverse the relation information back to the pixels. With the global structural and local consistency modules, these errors are rectified and confident pseudo-labels are generated for retraining. Extensive experiments on the LIP and the ATR datasets demonstrate the effectiveness of our global and local rectification modules. Our method outperforms other state-of-the-art methods in supervised human parsing tasks.
****

HigherHRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation
Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S. Huang, Lei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5386-5395
Bottom-up human pose estimation methods have difficulties in predicting the correct pose for small persons due to challenges in scale variation. In this paper, we present HigherHRNet: a novel bottom-up human pose estimation method for learning scale-aware representations using high-resolution feature pyramids. Equipped with multi-resolution supervision for training and multi-resolution aggregation for inference, the proposed approach is able to solve the scale variation challenge in bottom-up multi-person pose estimation and localize keypoints more precisely, especially for small person. The feature pyramid in HigherHRNet consists of feature map outputs from HRNet and upsampled higher-resolution outputs through a transposed convolution. HigherHRNet outperforms the previous best bottom-up method by 2.5% AP for medium person on COCO test-dev, showing its effectiveness in handling scale variation. Furthermore, HigherHRNet achieves new state-of-the-art result on COCO test-dev (70.5% AP) without using refinement or other post-processing techniques, surpassing all existing bottom-up methods. HigherHRNet even surpasses all top-down methods on CrowdPose test (67.6% AP), suggesting its robustness in crowded scene.
****

CNN-Generated Images Are Surprisingly Easy to Spot... for Now
Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, Alexei A. Efros; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8695-8704
In this work we ask whether it is possible to create a "universal" detector for telling apart real images from these generated by a CNN, regardless of architecture or dataset used. To test this, we collect a dataset consisting of fake images generated by 11 different CNN-based image generator models, chosen to span the space of commonly used architectures today (ProGAN, StyleGAN, BigGAN, CycleGAN, StarGAN, GauGAN, DeepFakes, cascaded refinement networks, implicit maximum likelihood estimation, second-order attention super-resolution, seeing-in-the-dark). We demonstrate that, with careful pre- and post-processing and data augmentation, a standard image classifier trained on only one specific CNN generator (ProGAN) is able to generalize surprisingly well to unseen architectures, datasets, and training methods (including the just released StyleGAN2). Our findings suggest the intriguing possibility that today's CNN-generated images share some common systematic flaws, preventing them from achieving realistic image synthesis.
****

Determinant Regularization for Gradient-Efficient Graph Matching
Tianshu Yu, Junchi Yan, Baoxin Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7123-7132

Graph matching refers to finding vertex correspondence for a pair of graphs, whi ch plays a fundamental role in many vision and learning related tasks. Directly applying gradient-based continuous optimization on graph matching can be attract ive for its simplicity but calls for effective ways of converting the continuous solution to the discrete one under the matching constraint. In this paper, we s how a novel regularization technique with the tool of determinant analysis on th e matching matrix which is relaxed into continuous domain with gradient based op timization. Meanwhile we present a theoretical study on the property of our rela xation technique. Our paper strikes an attempt to understand the geometric prope rties of different regularization techniques and the gradient behavior during th e optimization. We show that the proposed regularization is more gradient-effici ent than traditional ones during early update stages. The analysis will also bri ng about insights for other problems under bijection constraints. The algorithm procedure is simple and empirical results on public benchmark show its effective ness on both synthetic and real-world data.

****

A Stochastic Conditioning Scheme for Diverse Human Motion Prediction

Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, S tephen Gould; Proceedings of the IEEE/CVF Conference on Computer Vision and Patt ern Recognition (CVPR), 2020, pp. 5223-5232

Human motion prediction, the task of predicting future 3D human poses given a se quence of observed ones, has been mostly treated as a deterministic problem. How ever, human motion is a stochastic process: Given an observed sequence of poses, multiple future motions are plausible. Existing approaches to modeling this sto chasticity typically combine a random noise vector with information about the pr evious poses. This combination, however, is done in a deterministic manner, whic h gives the network the flexibility to learn to ignore the random noise. Alterna tively, in this paper, we propose to stochastically combine the root of variatio ns with previous pose information, so as to force the model to take the noise in to account. We exploit this idea for motion prediction by incorporating it into a recurrent encoder-decoder network with a conditional variational autoencoder b lock that learns to exploit the perturbations. Our experiments on two large-scal e motion prediction datasets demonstrate that our model yields high-quality pose sequences that are much more diverse than those from state-of-the-art stochasti c motion prediction techniques.

****

Can Facial Pose and Expression Be Separated With Weak Perspective Camera?

Evangelos Sariyanidi, Casey J. Zampella, Robert T. Schultz, Birkan Tunc; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2020, pp. 7173-7182

Separating facial pose and expression within images requires a camera model for 3D-to-2D mapping. The weak perspective (WP) camera has been the most popular cho ice; it is the default, if not the only option, in state-of-the-art facial analy sis methods and software. WP camera is justified by the supposition that its err ors are negligible when the subjects are relatively far from the camera, yet thi s claim has never been tested despite nearly 20 years of research. This paper cr itically examines the suitability of WP camera for separating facial pose and ex pression. First, we theoretically show that WP causes pose-expression ambiguity, as it leads to estimation of spurious expressions. Next, we experimentally quan tify the magnitude of spurious expressions. Finally, we test whether spurious ex pressions have detrimental effects on a common facial analysis application, name ly Action Unit (AU) detection. Contrary to conventional wisdom, we find that sev ere pose-expression ambiguity exists even when subjects are not close to the cam era, leading to large false positive rates in AU detection. We also demonstrate that the magnitude and characteristics of spurious expressions depend on the poi nt distribution model used to model the expressions. Our results suggest that co mmon assumptions about WP need to be revisited in facial expression modeling, an d that facial analysis software should encourage and facilitate the use of the t rue camera model whenever possible.

****

Probability Weighted Compact Feature for Domain Adaptive Retrieval
Fuxiang Huang, Lei Zhang, Yang Yang, Xichuan Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9582-9591
Domain adaptive image retrieval includes single-domain retrieval and cross-domain retrieval. Most of the existing image retrieval methods only focus on single-domain retrieval, which assumes that the distributions of retrieval databases and queries are similar. However, in practical application, the discrepancies between retrieval databases often taken in ideal illumination/pose/background/camera conditions and queries usually obtained in uncontrolled conditions are very large. In this paper, considering the practical application, we focus on challenging cross-domain retrieval. To address the problem, we propose an effective method named Probability Weighted Compact Feature Learning (PWCF), which provides inter-domain correlation guidance to promote cross-domain retrieval accuracy and learns a series of compact binary codes to improve the retrieval speed. First, we derive our loss function through the Maximum A Posteriori Estimation (MAP): Bayesian Perspective (BP) induced focal-triplet loss, BP induced quantization loss and BP induced classification loss. Second, we propose a common manifold structure between domains to explore the potential correlation across domains. Considering the original feature representation is biased due to the inter-domain discrepancy, the manifold structure is difficult to be constructed. Therefore, we propose a new feature named Histogram Feature of Neighbors (HFON) from the sample statistics perspective. Extensive experiments on various benchmark databases validate that our method outperforms many state-of-the-art image retrieval methods for domain adaptive image retrieval. The source code is available at https://github.com/fuxianghuang1/PWCF .
****
Compositional Convolutional Neural Networks: A Deep Architecture With Innate Robustness to Partial Occlusion
Adam Kortylewski, Ju He, Qing Liu, Alan L. Yuille; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8940-8949
Recent work has shown that deep convolutional neural networks (DCNNs) do not generalize well under partial occlusion. Inspired by the success of compositional models at classifying partially occluded objects, we propose to integrate compositional models and DCNNs into a unified deep model with innate robustness to partial occlusion. We term this architecture Compositional Convolutional Neural Network. In particular, we propose to replace the fully connected classification head of a DCNN with a differentiable compositional model. The generative nature of the compositional model enables it to localize occluders and subsequently focus on the non-occluded parts of the object. We conduct classification experiments on artificially occluded images as well as real images of partially occluded objects from the MS-COCO dataset. The results show that DCNNs do not classify occluded objects robustly, even when trained with data that is strongly augmented with partial occlusions. Our proposed model outperforms standard DCNNs by a large margin at classifying partially occluded objects, even when it has not been exposed to occluded objects during training. Additional experiments demonstrate that CompositionalNets can also localize the occluders accurately, despite being trained with class labels only. The code and data used in this work are publicly available.
****
Cascade EF-GAN: Progressive Facial Expression Editing With Local Focuses
Rongliang Wu, Gongjie Zhang, Shijian Lu, Tao Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5021-5030
Recent advances in Generative Adversarial Nets (GANs) have shown remarkable improvements for facial expression editing. However, current methods are still prone to generate artifacts and blurs around expression-intensive regions, and often introduce undesired overlapping artifacts while handling large-gap expression transformations such as transformation from furious to laughing. To address these

limitations, we propose Cascade Expression Focal GAN (Cascade EF-GAN), a novel n etwork that performs progressive facial expression editing with local expression focuses. The introduction of the local focus enables the Cascade EF-GAN to bett er preserve identity-related features and details around eyes, noses and mouths, which further helps reduce artifacts and blurs within the generated facial imag es. In addition, an innovative cascade transformation strategy is designed by di viding a large facial expression transformation into multiple small ones in casc ade, which helps suppress overlapping artifacts and produce more realistic editi ng while dealing with large-gap expression transformations. Extensive experiment s over two publicly available facial expression datasets show that our proposed Cascade EF-GAN achieves superior performance for facial expression editing.
****

TPNet: Trajectory Proposal Network for Motion Prediction
Liangji Fang, Qinhong Jiang, Jianping Shi, Bolei Zhou; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 67 97-6806
Making accurate motion prediction of the surrounding traffic agents such as pede strians, vehicles, and cyclists is crucial for autonomous driving. Recent data-d riven motion prediction methods have attempted to learn to directly regress the exact future position or its distribution from massive amount of trajectory data . However, it remains difficult for these methods to provide multimodal predicti ons as well as integrate physical constraints such as traffic rules and movable areas. In this work we propose a novel two-stage motion prediction framework, Tr ajectory Proposal Network (TPNet). TPNet first generates a candidate set of futu re trajectories as hypothesis proposals, then makes the final predictions by cla ssifying and refining the proposals which meets the physical constraints. By ste ering the proposal generation process, safe and multimodal predictions are reali zed. Thus this framework effectively mitigates the complexity of motion predicti on problem while ensuring the multimodal output. Experiments on four large-scale trajectory prediction datasets, i.e. the ETH, UCY, Apollo and Argoverse dataset s, show that TPNet achieves the state-of-the-art results both quantitatively and qualitatively.
****

Part-Aware Context Network for Human Parsing
Xiaomei Zhang, Yingying Chen, Bingke Zhu, Jinqiao Wang, Ming Tang; Proceedin gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8971-8980
Recent works have made significant progress in human parsing by exploiting rich contexts. However, human parsing still faces a challenge of how to generate adap tive contextual features for the various sizes and shapes of human parts. In thi s work, we propose a Part-aware Context Network (PCNet), a novel and effective a lgorithm to deal with the challenge. PCNet mainly consists of three modules, inc luding a part class module, a relational aggregation module, and a relational di spersion module. The part class module extracts the high-level representations o f every human part from a categorical perspective. We design a relational aggreg ation module to capture the representative global context by mining associated s emantics of human parts, which adaptively augments the context for human parts. We propose a relational dispersion module to generate the discriminative and eff ective local context and neglect disturbing one by making the affinity of human parts dispersed. The relational dispersion module ensures that features in the s ame class will be close to each other and away from those of different classes. By fusing the outputs of the relational aggregation module, the relational dispe rsion module and the backbone network, our PCNet generates adaptive contextual f eatures for various sizes of human parts, improving the parsing accuracy. We ach ieve a new state-of-the-art segmentation performance on three challenging human parsing datasets, i.e., PASCAL-Person-Part, LIP, and CIHP.
****

Lighthouse: Predicting Lighting Volumes for Spatially-Coherent Illumination
Pratul P. Srinivasan, Ben Mildenhall, Matthew Tancik, Jonathan T. Barron, Ri chard Tucker, Noah Snavely; Proceedings of the IEEE/CVF Conference on Computer

Vision and Pattern Recognition (CVPR), 2020, pp. 8080-8089
We present a deep learning solution for estimating the incident illumination at any 3D location within a scene from an input narrow-baseline stereo image pair. Previous approaches for predicting global illumination from images either predict just a single illumination for the entire scene, or separately estimate the illumination at each 3D location without enforcing that the predictions are consistent with the same 3D scene. Instead, we propose a deep learning model that estimates a 3D volumetric RGBA model of a scene, including content outside the observed field of view, and then uses standard volume rendering to estimate the incident illumination at any 3D location within that volume. Our model is trained without any ground truth 3D data and only requires a held-out perspective view near the input stereo pair and a spherical panorama taken within each scene as supervision, as opposed to prior methods for spatially-varying lighting estimation, which require ground truth scene geometry for training. We demonstrate that our method can predict consistent spatially-varying lighting that is convincing enough to plausibly relight and insert highly specular virtual objects into real images.
****

Joint Texture and Geometry Optimization for RGB-D Reconstruction
Yanping Fu, Qingan Yan, Jie Liao, Chunxia Xiao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5950-5959
Due to inevitable noises and quantization error, the reconstructed 3D models via RGB-D sensors always accompany geometric error and camera drifting, which consequently lead to blurring and unnatural texture mapping results. Most of the 3D reconstruction methods focus on either geometry refinement or texture improvement respectively, which subjectively decouples the inter-relationship between geometry and texture. In this paper, we propose a novel approach that can jointly optimize the camera poses, texture and geometry of the reconstructed model, and color consistency between the key-frames. Instead of computing Shape-From-Shading (SFS) expensively, our method directly optimizes the reconstructed mesh according to color and geometric consistency and high-boost normal cues, which can effectively overcome the texture-copy problem generated by SFS and achieve more detailed shape reconstruction. As the joint optimization involves multiple correlated terms, therefore, we further introduce an iterative framework to interleave the optimal state. The experiments demonstrate that our method can recover not only fine-scale geometry but also high-fidelity texture.
****

Hyperbolic Visual Embedding Learning for Zero-Shot Recognition
Shaoteng Liu, Jingjing Chen, Liangming Pan, Chong-Wah Ngo, Tat-Seng Chua, Yu-Gang Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9273-9281
This paper proposes a Hyperbolic Visual Embedding Learning Network for zero-shot recognition. The network learns image embeddings in hyperbolic space, which is capable of preserving the hierarchical structure of semantic classes in low dimensions. Comparing with existing zero-shot learning approaches, the network is more robust because the embedding feature in hyperbolic space better represents class hierarchy and thereby avoid misleading resulted from unrelated siblings. Our network outperforms exiting baselines under hierarchical evaluation with an extremely challenging setting, i.e., learning only from 1,000 categories to recognize 20,841 unseen categories. While under flat evaluation, it has competitive performance as state-of-the-art methods but with five times lower embedding dimensions. Our code is publicly available (https://github.com/ShaoTengLiu/Hyperbolic_ZSL).
****

LSM: Learning Subspace Minimization for Low-Level Vision
Chengzhou Tang, Lu Yuan, Ping Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6235-6246
We study the energy minimization problem in low-level vision tasks from a novel perspective. We replace the heuristic regularization term with a data-driven learnable subspace constraint, and preserve the data term to exploit domain knowled

ge derived from the first principles of a task. This learning subspace minimization (LSM) framework unifies the network structures and the parameters for many different low-level vision tasks, which allows us to train a single network for multiple tasks simultaneously with shared parameters, and even generalizes the trained network to an unseen task as long as the data term can be formulated. We validate our LSM frame on four low-level tasks including edge detection, interactive segmentation, stereo matching, and optical flow, and validate the network on various datasets. The experiments demonstrate that the proposed LSM generates state-of-the-art results with smaller model size, faster training convergence, and real-time inference.

****

Erasing Integrated Learning: A Simple Yet Effective Approach for Weakly Supervised Object Localization

Jinjie Mai, Meng Yang, Wenfeng Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8766-8775

Weakly supervised object localization (WSOL) aims to localize object with only weak supervision like image-level labels. However, a long-standing problem for available techniques based on the classification network is that they often result in highlighting the most discriminative parts rather than the entire extent of object. Nevertheless, trying to explore the integral extent of the object could degrade the performance of image classification on the contrary. To remedy this, we propose a simple yet powerful approach by introducing a novel adversarial erasing technique, erasing integrated learning (EIL). By integrating discriminative region mining and adversarial erasing in a single forward-backward propagation in a vanilla CNN, the proposed EIL explores the high response class-specific area and the less discriminative region simultaneously, thus could maintain high performance in classification and jointly discover the full extent of the object. Furthermore, we apply multiple EIL (MEIL) modules at different levels of the network in a sequential manner, which for the first time integrates semantic features of multiple levels and multiple scales through adversarial erasing learning. In particular, the proposed EIL and advanced MEIL both achieve a new state-of-the-art performance in CUB-200-2011 and ILSVRC 2016 benchmark, making significant improvement in localization while advancing high performance in image classification.

****

Self-Supervised Deep Visual Odometry With Online Adaptation

Shunkai Li, Xin Wang, Yingdian Cao, Fei Xue, Zike Yan, Hongbin Zha; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6339-6348

Self-supervised VO methods have shown great success in jointly estimating camera pose and depth from videos. However, like most data-driven methods, existing VO networks suffer from a notable decrease in performance when confronted with scenes different from the training data, which makes them unsuitable for practical applications. In this paper, we propose an online meta-learning algorithm to enable VO networks to continuously adapt to new environments in a self-supervised manner. The proposed method utilizes convolutional long short-term memory (convLSTM) to aggregate rich spatial-temporal information in the past. The network is able to memorize and learn from its past experience for better estimation and fast adaptation to the current frame. When running VO in the open world, in order to deal with the changing environment, we propose an online feature alignment method by aligning feature distributions at different time. Our VO network is able to seamlessly adapt to different environments. Extensive experiments on unseen outdoor scenes, virtual to real world and outdoor to indoor environments demonstrate that our method consistently outperforms state-of-the-art self-supervised VO baselines considerably.

****

Weakly-Supervised Semantic Segmentation via Sub-Category Exploration

Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, Ming-Hsuan Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8991-9000

Existing weakly-supervised semantic segmentation methods using image-level annotations typically rely on initial responses to locate object regions. However, such response maps generated by the classification network usually focus on discriminative object parts, due to the fact that the network does not need the entire object for optimizing the objective function. To enforce the network to pay attention to other parts of an object, we propose a simple yet effective approach that introduces a self-supervised task by exploiting the sub-category information. Specifically, we perform clustering on image features to generate pseudo sub-categories labels within each annotated parent class, and construct a sub-category objective to assign the network to a more challenging task. By iteratively clustering image features, the training process does not limit itself to the most discriminative object parts, hence improving the quality of the response maps. We conduct extensive analysis to validate the proposed method and show that our approach performs favorably against the state-of-the-art approaches.
****

Normalizing Flows With Multi-Scale Autoregressive Priors
Apratim Bhattacharyya, Shweta Mahajan, Mario Fritz, Bernt Schiele, Stefan Roth; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8415-8424
Flow-based generative models are an important class of exact inference models that admit efficient inference and sampling for image synthesis. Owing to the efficiency constraints on the design of the flow layers, e.g. split coupling flow layers in which approximately half the pixels do not undergo further transformations, they have limited expressiveness for modeling long-range data dependencies compared to autoregressive models that rely on conditional pixel-wise generation. In this work, we improve the representational power of flow-based models by introducing channel-wise dependencies in their latent space through multi-scale autoregressive priors (mAR). Our mAR prior for models with split coupling flow layers (mAR-SCF) can better capture dependencies in complex multimodal data. The resulting model achieves state-of-the-art density estimation results on MNIST, CIFAR-10, and ImageNet. Furthermore, we show that mAR-SCF allows for improved image generation quality, with gains in FID and Inception scores compared to state-of-the-art flow-based models.
****

Dynamic Neural Relational Inference
Colin Graber, Alexander G. Schwing; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8513-8522
Understanding interactions between entities, e.g., joints of the human body, team sports players, etc., is crucial for tasks like forecasting. However, interactions between entities are commonly not observed and often hard to quantify. To address this challenge, recently, `Neural Relational Inference' was introduced. It predicts static relations between entities in a system and provides an interpretable representation of the underlying system dynamics that are used for better trajectory forecasting. However, generally, relations between entities change as time progresses. Hence, static relations improperly model the data. In response to this, we develop Dynamic Neural Relational Inference (dNRI), which incorporates insights from sequential latent variable models to predict separate relation graphs for every time-step. We demonstrate on several real-world datasets that modeling dynamic relations improves forecasting of complex trajectories.
****

Embedding Expansion: Augmentation in Embedding Space for Deep Metric Learning
Byungsoo Ko, Geonmo Gu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7255-7264
Learning the distance metric between pairs of samples has been studied for image retrieval and clustering. With the remarkable success of pair-based metric learning losses, recent works have proposed the use of generated synthetic points on metric learning losses for augmentation and generalization. However, these methods require additional generative networks along with the main network, which can lead to a larger model size, slower training speed, and harder optimization. Meanwhile, post-processing techniques, such as query expansion and database augme

ntation, have proposed the combination of feature points to obtain additional se
mantic information. In this paper, inspired by query expansion and database augm
entation, we propose an augmentation method in an embedding space for pair-based
 metric learning losses, called embedding expansion. The proposed method generat
es synthetic points containing augmented information by a combination of feature
 points and performs hard negative pair mining to learn with the most informativ
e feature representations. Because of its simplicity and flexibility, it can be
used for existing metric learning losses without affecting model size, training
speed, or optimization difficulty. Finally, the combination of embedding expansi
on and representative metric learning losses outperforms the state-of-the-art lo
sses and previous sample generation methods in both image retrieval and clusteri
ng tasks. The implementation is publicly available.
****

LT-Net: Label Transfer by Learning Reversible Voxel-Wise Correspondence for One-Shot Medical Image Segmentation

Shuxin Wang, Shilei Cao, Dong Wei, Renzhen Wang, Kai Ma, Liansheng Wang, Deyu Meng, Yefeng Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9162-9171

We introduce a one-shot segmentation method to alleviate the burden of manual an
notation for medical images. The main idea is to treat one-shot segmentation as
a classical atlas-based segmentation problem, where voxel-wise correspondence fr
om the atlas to the unlabelled data is learned. Subsequently, segmentation label
 of the atlas can be transferred to the unlabelled data with the learned corresp
ondence. However, since ground truth correspondence between images is usually un
available, the learning system must be well-supervised to avoid mode collapse an
d convergence failure. To overcome this difficulty, we resort to the forward-bac
kward consistency, which is widely used in correspondence problems, and addition
ally learn the backward correspondences from the warped atlases back to the orig
inal atlas. This cycle-correspondence learning design enables a variety of extra
, cycle-consistency-based supervision signals to make the training process stabl
e, while also boost the performance. We demonstrate the superiority of our metho
d over both deep learning-based one-shot segmentation methods and a classical mu
lti-atlas segmentation method via thorough experiments.
****

Transferring Dense Pose to Proximal Animal Classes

Artsiom Sanakoyeu, Vasil Khalidov, Maureen S. McCarthy, Andrea Vedaldi, Natalia Neverova; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5233-5242

Recent contributions have demonstrated that it is possible to recognize the pose
 of humans densely and accurately given a large dataset of poses annotated in de
tail. In principle, the same approach could be extended to any animal class, but
 the effort required for collecting new annotations for each case makes this str
ategy impractical, despite important applications in natural conservation, scien
ce and business. We show that, at least for proximal animal classes such as chim
panzees, it is possible to transfer the knowledge existing in dense pose recogni
tion for humans, as well as in more general object detectors and segmenters, to
the problem of dense pose recognition in other classes. We do this by (1) establ
ishing a DensePose model for the new animal which is also geometrically aligned
to humans (2) introducing a multi-head R-CNN architecture that facilitates trans
fer of multiple recognition tasks between classes, (3) finding which combination
 of known classes can be transferred most effectively to the new animal and (4)
using self-calibrated uncertainty heads to generate pseudo-labels graded by qual
ity for training a model for this class. We also introduce two benchmark dataset
s labelled in the manner of DensePose for the class chimpanzee and use them to e
valuate our approach, showing excellent transfer learning performance.
****

Suppressing Uncertainties for Large-Scale Facial Expression Recognition

Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, Yu Qiao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6897-6906

Annotating a qualitative large-scale facial expression dataset is extremely difficult due to the uncertainties caused by ambiguous facial expressions, low-quality facial images, and the subjectiveness of annotators. These uncertainties suspend the progress of large-scale Facial Expression Recognition (FER) in data-driven deep learning era. To address this problelm, this paper proposes to suppress the uncertainties by a simple yet efficient Self-Cure Network (SCN). Specifically, SCN suppresses the uncertainty from two different aspects: 1) a self-attention mechanism over FER dataset to weight each sample in training with a ranking regularization, and 2) a careful relabeling mechanism to modify the labels of these samples in the lowest-ranked group. Experiments on synthetic FER datasets and our collected WebEmotion dataset validate the effectiveness of our method. Results on public benchmarks demonstrate that our SCN outperforms current state-of-the-art methods with 88.14% on RAF-DB, 60.23% on AffectNet, and 89.35% on FERPlus.
****

Scale-Space Flow for End-to-End Optimized Video Compression
Eirikur Agustsson, David Minnen, Nick Johnston, Johannes Balle, Sung Jin Hwang, George Toderici; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8503-8512
Despite considerable progress on end-to-end optimized deep networks for image compression, video coding remains a challenging task. Recently proposed methods for learned video compression use optical flow and bilinear warping for motion compensation and show competitive rate-distortion performance relative to hand-engineered codecs like H.264 and HEVC. However, these learning-based methods rely on complex architectures and training schemes including the use of pre-trained optical flow networks, sequential training of sub-networks, adaptive rate control, and buffering intermediate reconstructions to disk during training. In this paper, we show that a generalized warping operator that better handles common failure cases, e.g. disocclusions and fast motion, can provide competitive compression results with a greatly simplified model and training procedure. Specifically, we propose scale-space flow, an intuitive generalization of optical flow that adds a scale parameter to allow the network to better model uncertainty. Our experiments show that a low-latency video compression model (no B-frames) using scale-space flow for motion compensation can outperform analogous state-of-the art learned video compression models while being trained using a much simpler procedure and without any pre-trained optical flow networks.
****

StyleRig: Rigging StyleGAN for 3D Control Over Portrait Images
Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Perez, Michael Zollhofer, Christian Theobalt; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6142-6151
StyleGAN generates photorealistic portrait images of faces with eyes, teeth, hair and context (neck, shoulders, background), but lacks a rig-like control over semantic face parameters that are interpretable in 3D, such as face pose, expressions, and scene illumination. Three-dimensional morphable face models (3DMMs) on the other hand offer control over the semantic parameters, but lack photorealism when rendered and only model the face interior, not other parts of a portrait image (hair, mouth interior, background). We present the first method to provide a face rig-like control over a pretrained and fixed StyleGAN via a 3DMM. A new rigging network, RigNet is trained between the 3DMM's semantic parameters and StyleGAN's input. The network is trained in a self-supervised manner, without the need for manual annotations. At test time, our method generates portrait images with the photorealism of StyleGAN and provides explicit control over the 3D semantic parameters of the face.
****

Semantic Pyramid for Image Generation
Assaf Shocher, Yossi Gandelsman, Inbar Mosseri, Michal Yarom, Michal Irani, William T. Freeman, Tali Dekel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7457-7466
We present a novel GAN-based model that utilizes the space of deep features lear

ned by a pre-trained classification model. Inspired by classical image pyramid r
epresentations, we construct our model as a Semantic Generation Pyramid -- a hie
rarchical framework which leverages the continuum of semantic information encaps
ulated in such deep features; this ranges from low level information contained i
n fine features to high level, semantic information contained in deeper features
. More specifically, given a set of features extracted from a reference image, o
ur model generates diverse image samples, each with matching features at each se
mantic level of the classification model. We demonstrate that our model results
in a versatile and flexible framework that can be used in various classic and no
vel image generation tasks. These include: generating images with a controllable
 extent of semantic similarity to a reference image, and different manipulation
tasks such as semantically-controlled inpainting and compositing; all achieved w
ith the same model, with no further training.
****

Towards Backward-Compatible Representation Learning
Yantao Shen, Yuanjun Xiong, Wei Xia, Stefano Soatto; Proceedings of the IEEE/
CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6368
-6377
We propose a way to learn visual features that are compatible with previously co
mputed ones even when they have different dimensions and are learned via differe
nt neural network architectures and loss functions. Compatible means that, if su
ch features are used to compare images, then "new" features can be compared dire
ctly to "old" features, so they can be used interchangeably. This enables visual
 search systems to bypass computing new features for all previously seen images
when updating the embedding models, a process known as backfilling. Backward com
patibility is critical to quickly deploy new embedding models that leverage ever
-growing large-scale training datasets and improvements in deep learning archite
ctures and training methods. We propose a framework to train embedding models, c
alled backward-compatible training (BCT), as a first step towards backward compa
tible representation learning. In experiments on learning embeddings for face re
cognition, models trained with BCT successfully achieve backward compatibility w
ithout sacrificing accuracy, thus enabling backfill-free model updates of visual
 embeddings.
****

Global-Local GCN: Large-Scale Label Noise Cleansing for Face Recognition
Yaobin Zhang, Weihong Deng, Mei Wang, Jiani Hu, Xian Li, Dongyue Zhao, Don
gchao Wen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
 Recognition (CVPR), 2020, pp. 7731-7740
In the field of face recognition, large-scale web-collected datasets are essenti
al for learning discriminative representations, but they suffer from noisy ident
ity labels, such as outliers and label flips. It is beneficial to automatically
cleanse their label noise for improving recognition accuracy. Unfortunately, exi
sting cleansing methods cannot accurately identify noise in the wild. To solve t
his problem, we propose an effective automatic label noise cleansing framework f
or face recognition datasets, FaceGraph. Using two cascaded graph convolutional
networks, FaceGraph performs global-to-local discrimination to select useful dat
a in a noisy environment. Extensive experiments show that cleansing widely used
datasets, such as CASIA-WebFace, VGGFace2, MegaFace2, and MS-Celeb-1M, using the
 proposed method can improve the recognition performance of state-of-the-art rep
resentation learning methods like Arcface. Further, we cleanse massive self-coll
ected celebrity data, namely MillionCelebs, to provide 18.8M images of 636K iden
tities. Training with the new data, Arcface surpasses state-of-the-art performan
ce by a notable margin to reach 95.62% TPR at 1e-5 FPR on the IJB-C benchmark.
****

Adaptive Graph Convolutional Network With Attention Graph Clustering for Co-Sali
ency Detection
Kaihua Zhang, Tengpeng Li, Shiwen Shen, Bo Liu, Jin Chen, Qingshan Liu; Pro
ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (
CVPR), 2020, pp. 9050-9059
Co-saliency detection aims to discover the common and salient foregrounds from a

group of relevant images. For this task, we present a novel adaptive graph convolutional network with attention graph clustering (GCAGC). Three major contributions have been made, and are experimentally shown to have substantial practical merits. First, we propose a graph convolutional network design to extract information cues to characterize the intra- and inter-image correspondence. Second, we develop an attention graph clustering algorithm to discriminate the common objects from all the salient foreground objects in an unsupervised fashion. Third, we present a unified framework with encoder-decoder structure to jointly train and optimize the graph convolutional network, attention graph cluster, and co-saliency detection decoder in an end-to-end manner. We evaluate our proposed GCAGC method on three co-saliency detection benchmark datasets (iCoseg, Cosal2015 and COCO-SEG). Our GCAGC method obtains significant improvements over the state-of-the-arts on most of them.

****

UniPose: Unified Human Pose Estimation in Single Images and Videos

Bruno Artacho, Andreas Savakis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7035-7044

We propose UniPose, a unified framework for human pose estimation, based on our "Waterfall" Atrous Spatial Pooling architecture, that achieves state-of-art-results on several pose estimation metrics. UniPose incorporates contextual segmentation and joint localization to estimate the human pose in a single stage, with high accuracy, without relying on statistical postprocessing methods. The Waterfall module in UniPose leverages the efficiency of progressive filtering in the cascade architecture, while maintaining multi-scale fields-of-view comparable to spatial pyramid configurations. Additionally, our method is extended to UniPose-LSTM for multi-frame processing and achieves state-of-the-art results for temporal pose estimation in Video. Our results on multiple datasets demonstrate that UniPose, with a ResNet backbone and Waterfall module, is a robust and efficient architecture for pose estimation obtaining state-of-the-art results in single person pose detection for both single images and videos.

****

Novel View Synthesis of Dynamic Scenes With Globally Coherent Depths From a Monocular Camera

Jae Shin Yoon, Kihwan Kim, Orazio Gallo, Hyun Soo Park, Jan Kautz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5336-5345

This paper presents a new method to synthesize an image from arbitrary views and times given a collection of images of a dynamic scene. A key challenge for the novel view synthesis arises from dynamic scene reconstruction where epipolar geometry does not apply to the local motion of dynamic contents. To address this challenge, we propose to combine the depth from single view (DSV) and the depth from multi-view stereo (DMV), where DSV is complete, i.e., a depth is assigned to every pixel, yet view-variant in its scale, while DMV is view-invariant yet incomplete. Our insight is that although its scale and quality are inconsistent with other views, the depth estimation from a single view can be used to reason about the globally coherent geometry of dynamic contents. We cast this problem as learning to correct the scale of DSV, and to refine each depth with locally consistent motions between views to form a coherent depth estimation. We integrate these tasks into a depth fusion network in a self-supervised fashion. Given the fused depth maps, we synthesize a photorealistic virtual view in a specific location and time with our deep blending network that completes the scene and renders the virtual view. We evaluate our method of depth estimation and view synthesis on a diverse real-world dynamic scenes and show the outstanding performance over existing methods.

****

Cogradient Descent for Bilinear Optimization

Li'an Zhuo, Baochang Zhang, Linlin Yang, Hanlin Chen, Qixiang Ye, David Doermann, Rongrong Ji, Guodong Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7959-7967

Conventional learning methods simplify the bilinear model by regarding two intri

nsically coupled factors independently, which degrades the optimization procedur e. One reason lies in the insufficient training due to the asynchronous gradient descent, which results in vanishing gradients for the coupled variables. In thi s paper, we introduce a Cogradient Descent algorithm (CoGD) to address the bilin ear problem, based on a theoretical framework to coordinate the gradient of hidd en variables via a projection function. We solve one variable by considering its coupling relationship with the other, leading to a synchronous gradient descent to facilitate the optimization procedure. Our algorithm is applied to solve pro blems with one variable under the sparsity constraint, which is widely used in t he learning paradigm. We validate our CoGD considering an extensive set of appli cations including image reconstruction, inpainting, and network pruning. Experim ents show that it improves the state-of-the-art by a significant margin.
****

AdversarialNAS: Adversarial Neural Architecture Search for GANs
Chen Gao, Yunpeng Chen, Si Liu, Zhenxiong Tan, Shuicheng Yan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020 , pp. 5680-5689
Neural Architecture Search (NAS) that aims to automate the procedure of architec ture design has achieved promising results in many computer vision fields. In th is paper, we propose an AdversarialNAS method specially tailored for Generative Adversarial Networks (GANs) to search for a superior generative model on the tas k of unconditional image generation. The AdversarialNAS is the first method that can search the architectures of generator and discriminator simultaneously in a differentiable manner. During searching, the designed adversarial search algori thm does not need to comput any extra metric to evaluate the performance of the searched architecture, and the search paradigm considers the relevance between t he two network architectures and improves their mutual balance. Therefore, Adver sarialNAS is very efficient and only takes 1 GPU day to search for a superior ge nerative model in the proposed large search space. Experiments demonstrate the e ffectiveness and superiority of our method. The discovered generative model sets a new state-of-the-art FID score of 10.87 and highly competitive Inception Scor e of 8.74 on CIFAR-10. Its transferability is also proven by setting new state-o f-the-art FID score of 26.98 and Inception score of 9.63 on STL-10. Code is at: https://github.com/chengaopro/AdversarialNAS.
****

Belief Propagation Reloaded: Learning BP-Layers for Labeling Problems
Patrick Knobelreiter, Christian Sormann, Alexander Shekhovtsov, Friedrich Fra undorfer, Thomas Pock; Proceedings of the IEEE/CVF Conference on Computer Visio n and Pattern Recognition (CVPR), 2020, pp. 7900-7909
It has been proposed by many researchers that combining deep neural networks wit h graphical models can create more efficient and better regularized composite mo dels. The main difficulties in implementing this in practice are associated with a discrepancy in suitable learning objectives as well as with the necessity of approximations for the inference. In this work we take one of the simplest infer ence methods, a truncated max-product Belief Propagation, and add what is necess ary to make it a proper component of a deep learning model: connect it to learni ng formulations with losses on marginals and compute the backprop operation. Thi s BP-Layer can be used as the final or an intermediate block in convolutional ne ural networks (CNNs), allowing us to design a hierarchical model composing BP in ference and CNNs at different scale levels. The model is applicable to a range o f dense prediction problems, is well-trainable and provides parameter-efficient and robust solutions in stereo, flow and semantic segmentation.
****

DoveNet: Deep Image Harmonization via Domain Verification
Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, Liqin g Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern R ecognition (CVPR), 2020, pp. 8394-8403
Image composition is an important operation in image processing, but the inconsi stency between foreground and background significantly degrades the quality of c omposite image. Image harmonization, aiming to make the foreground compatible wi

th the background, is a promising yet challenging task. However, the lack of hig h-quality publicly available dataset for image harmonization greatly hinders the development of image harmonization techniques. In this work, we contribute an i mage harmonization dataset iHarmony4 by generating synthesized composite images based on COCO (resp., Adobe5k, Flickr, day2night) dataset, leading to our HCOCO (resp., HAdobe5k, HFlickr, Hday2night) sub-dataset. Moreover, we propose a new d eep image harmonization method DoveNet using a novel domain verification discrim inator, with the insight that the foreground needs to be translated to the same domain as background. Extensive experiments on our constructed dataset demonstra te the effectiveness of our proposed method. Our dataset and code are available at https://github.com/bcmi/Image_Harmonization_Datasets.
****

Self-Supervised 3D Human Pose Estimation via Part Guided Novel Image Synthesis
Jogendra Nath Kundu, Siddharth Seth, Varun Jampani, Mugalodi Rakesh, R. Venk atesh Babu, Anirban Chakraborty; Proceedings of the IEEE/CVF Conference on Comp uter Vision and Pattern Recognition (CVPR), 2020, pp. 6152-6162
Camera captured human pose is an outcome of several sources of variation. Perfor mance of supervised 3D pose estimation approaches comes at the cost of dispensin g with variations, such as shape and appearance, that may be useful for solving other related tasks. As a result, the learned model not only inculcates task-bia s but also dataset-bias because of its strong reliance on the annotated samples, which also holds true for weakly-supervised models. Acknowledging this, we prop ose a self-supervised learning framework to disentangle such variations from unl abeled video frames. We leverage the prior knowledge on human skeleton and poses in the form of a single part-based 2D puppet model, human pose articulation con straints, and a set of unpaired 3D poses. Our differentiable formalization, brid ging the representation gap between the 3D pose and spatial part maps, not only facilitates discovery of interpretable pose disentanglement, but also allows us to operate on videos with diverse camera movements. Qualitative results on unsee n in-the-wild datasets establish our superior generalization across multiple tas ks beyond the primary tasks of 3D pose estimation and part segmentation. Further more, we demonstrate state-of-the-art weakly-supervised 3D pose estimation perfo rmance on both Human3.6M and MPI-INF-3DHP datasets.
****

Self-Supervised Learning of Interpretable Keypoints From Unlabelled Videos
Tomas Jakab, Ankush Gupta, Hakan Bilen, Andrea Vedaldi; Proceedings of the IE EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8 787-8797
We propose a new method for recognizing the pose of objects from a single image that for learning uses only unlabelled videos and a weak empirical prior on the object poses. Video frames differ primarily in the pose of the objects they cont ain, so our method distils the pose information by analyzing the differences bet ween frames. The distillation uses a new dual representation of the geometry of objects as a set of 2D keypoints, and as a pictorial representation, i.e. a skel eton image. This has three benefits: (1) it provides a tight 'geometric bottlene ck' which disentangles pose from appearance, (2) it can leverage powerful image- to-image translation networks to map between photometry and geometry, and (3) it allows to incorporate empirical pose priors in the learning process. The pose p riors are obtained from unpaired data, such as from a different dataset or modal ity such as mocap, such that no annotated image is ever used in learning the pos e recognition network. In standard benchmarks for pose recognition for humans an d faces, our method achieves state-of-the-art performance among methods that do not require any labelled images for training. Project page: http://www.robots.ox .ac.uk/ vgg/research/unsupervised_pose/
****

Distribution-Aware Coordinate Representation for Human Pose Estimation
Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, Ce Zhu; Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 709 3-7102
While being the de facto standard coordinate representation for human pose estim

ation, heatmap has not been investigated in-depth. This work fills this gap. For the first time, we find that the process of decoding the predicted heatmaps into the final joint coordinates in the original image space is surprisingly significant for the performance. We further probe the design limitations of the standard coordinate decoding method, and propose a more principled distributionaware decoding method. Also, we improve the standard coordinate encoding process (i.e. transforming ground-truth coordinates to heatmaps) by generating unbiased/accurate heatmaps. Taking the two together, we formulate a novel Distribution-Aware coordinate Representation of Keypoints (DARK) method. Serving as a model-agnostic plug-in, DARK brings about significant performance boost to existing human pose estimation models. Extensive experiments show that DARK yields the best results on two common benchmarks, MPII and COCO. Besides, DARK achieves the 2nd place entry in the ICCV 2019 COCO Keypoints Challenge. The code is available online.
****

Attention Mechanism Exploits Temporal Contexts: Real-Time 3D Human Pose Reconstruction
Ruixu Liu, Ju Shen, He Wang, Chen Chen, Sen-ching Cheung, Vijayan Asari; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5064-5073
We propose a novel attention-based framework for 3D human pose estimation from a monocular video. Despite the general success of end-to-end deep learning paradigms, our approach is based on two key observations: (1) temporal incoherence and jitter are often yielded from a single frame prediction; (2) error rate can be remarkably reduced by increasing the receptive field in a video. Therefore, we design an attentional mechanism to adaptively identify significant frames and tensor outputs from each deep neural net layer, leading to a more optimal estimation. To achieve large temporal receptive fields, multi-scale dilated convolutions are employed to model long-range dependencies among frames. The architecture is straightforward to implement and can be flexibly adopted for real-time applications. Any off-the-shelf 2D pose estimation system, e.g. Mocap libraries, can be easily integrated in an ad-hoc fashion. We both quantitatively and qualitatively evaluate our method on various standard benchmark datasets (e.g. Human3.6M, HumanEva). Our method considerably outperforms all the state-of-the-art algorithms up to 8% error reduction (average mean per joint position error: 34.7) as compared to the best-reported results. Code is available at: (https://github.com/lrxjason/Attention3DHumanPose)
****

MaskFlownet: Asymmetric Feature Matching With Learnable Occlusion Mask
Shengyu Zhao, Yilun Sheng, Yue Dong, Eric I-Chao Chang, Yan Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6278-6287
Feature warping is a core technique in optical flow estimation; however, the ambiguity caused by occluded areas during warping is a major problem that remains unsolved. In this paper, we propose an asymmetric occlusion-aware feature matching module, which can learn a rough occlusion mask that filters useless (occluded) areas immediately after feature warping without any explicit supervision. The proposed module can be easily integrated into end-to-end network architectures and enjoys performance gains while introducing negligible computational cost. The learned occlusion mask can be further fed into a subsequent network cascade with dual feature pyramids with which we achieve state-of-the-art performance. At the time of submission, our method, called MaskFlownet, surpasses all published optical flow methods on the MPI Sintel, KITTI 2012 and 2015 benchmarks. Code is available at https://github.com/microsoft/MaskFlownet.
****

3FabRec: Fast Few-Shot Face Alignment by Reconstruction
Bjorn Browatzki, Christian Wallraven; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6110-6120
Current supervised methods for facial landmark detection require a large amount of training data and may suffer from overfitting to specific datasets due to the massive number of parameters. We introduce a semi-supervised method in which th

e crucial idea is to first generate implicit face knowledge from the large amoun ts of unlabeled images of faces available today. In a first, completely unsuperv ised stage, we train an adversarial autoencoder to reconstruct faces via a low-d imensional face embedding. In a second, supervised stage, we interleave the deco der with transfer layers to retask the generation of color images to the predict ion of landmark heatmaps. Our framework (3FabRec) achieves state-of-the-art perf ormance on several common benchmarks and, most importantly, is able to maintain impressive accuracy on extremely small training sets down to as few as 10 images . As the interleaved layers only add a low amount of parameters to the decoder, inference runs at several hundred FPS on a GPU.
****

MARMVS: Matching Ambiguity Reduced Multiple View Stereo for Efficient Large Scal e Scene Reconstruction
Zhenyu Xu, Yiguang Liu, Xuelei Shi, Ying Wang, Yunan Zheng; Proceedings of t he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5981-5990
The ambiguity in image matching is one of main factors decreasing the quality of the 3D model reconstructed by PatchMatch based multiple view stereo. In this pa per, we present a novel method, matching ambiguity reduced multiple view stereo (MARMVS) to address this issue. The MARMVS handles the ambiguity in image matchi ng process with three newly proposed strategies: 1) The matching ambiguity is me asured by the differential geometry property of image surface with epipolar cons traint, which is used as a critical criterion for optimal scale selection of eve ry single pixel with corresponding neighbouring images. 2) The depth of every pi xel is initialized to be more close to the true depth by utilizing the depths of its surrounding sparse feature points, which yields faster convergency speed in the following PatchMatch stereo and alleviates the ambiguity introduced by self similar structures of the image. 3) In the last propagation of the PatchMatch s tereo, higher priorities are given to those planes with the related 2D image pat ch possesses less ambiguity, this strategy further propagates a correctly recons tructed surface to raw texture regions. In addition, the proposed method is very efficient even running on consumer grade CPUs, due to proper parameterization a nd discretization in the depth map computation step. The MARMVS is validated on public benchmarks, and experimental results demonstrate competing performance ag ainst the state of the art.
****

Bodies at Rest: 3D Human Pose and Shape Estimation From a Pressure Image Using S ynthetic Data
Henry M. Clever, Zackory Erickson, Ariel Kapusta, Greg Turk, Karen Liu, Cha rles C. Kemp; Proceedings of the IEEE/CVF Conference on Computer Vision and Patt ern Recognition (CVPR), 2020, pp. 6215-6224
People spend a substantial part of their lives at rest in bed. 3D human pose and shape estimation for this activity would have numerous beneficial applications, yet line-of-sight perception is complicated by occlusion from bedding. Pressure sensing mats are a promising alternative, but training data is challenging to c ollect at scale. We describe a physics-based method that simulates human bodies at rest in a bed with a pressure sensing mat, and present PressurePose, a synthe tic dataset with 206K pressure images with 3D human poses and shapes. We also pr esent PressureNet, a deep learning model that estimates human pose and shape giv en a pressure image and gender. PressureNet incorporates a pressure map reconstr uction (PMR) network that models pressure image generation to promote consistenc y between estimated 3D body models and pressure image input. In our evaluations, PressureNet performed well with real data from participants in diverse poses, e ven though it had only been trained with synthetic data. When we ablated the PMR network, performance dropped substantially.
****

Cars Can't Fly Up in the Sky: Improving Urban-Scene Segmentation via Height-Driv en Attention Networks
Sungha Choi, Joanne T. Kim, Jaegul Choo; Proceedings of the IEEE/CVF Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9373-9383

This paper exploits the intrinsic features of urban-scene images and proposes a general add-on module, called height-driven attention networks (HANet), for improving semantic segmentation for urban-scene images. It emphasizes informative features or classes selectively according to the vertical position of a pixel. The pixel-wise class distributions are significantly different from each other among horizontally segmented sections in the urban-scene images. Likewise, urban-scene images have their own distinct characteristics, but most semantic segmentation networks do not reflect such unique attributes in the architecture. The proposed network architecture incorporates the capability exploiting the attributes to handle the urban scene dataset effectively. We validate the consistent performance (mIoU) increase of various semantic segmentation models on two datasets when HANet is adopted. This extensive quantitative analysis demonstrates that adding our module to existing models is easy and cost-effective. Our method achieves a new state-of-the-art performance on the Cityscapes benchmark with a large margin among ResNet101 based segmentation models. Also, we show that the proposed model is coherent with the facts observed in the urban scene by visualizing and interpreting the attention map. Our code and trained models are publicly available.
****

Compressed Volumetric Heatmaps for Multi-Person 3D Pose Estimation
Matteo Fabbri, Fabio Lanzi, Simone Calderara, Stefano Alletto, Rita Cucchiara; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7204-7213
In this paper we present a novel approach for bottom-up multi-person 3D human pose estimation from monocular RGB images. We propose to use high resolution volumetric heatmaps to model joint locations, devising a simple and effective compression method to drastically reduce the size of this representation. At the core of the proposed method lies our Volumetric Heatmap Autoencoder, a fully-convolutional network tasked with the compression of ground-truth heatmaps into a dense intermediate representation. A second model, the Code Predictor, is then trained to predict these codes, which can be decompressed at test time to re-obtain the original representation. Our experimental evaluation shows that our method performs favorably when compared to state of the art on both multi-person and single-person 3D human pose estimation datasets and, thanks to our novel compression strategy, can process full-HD images at the constant runtime of 8 fps regardless of the number of subjects in the scene. Code and models are publicly available.
****

3D-MPA: Multi-Proposal Aggregation for 3D Semantic Instance Segmentation
Francis Engelmann, Martin Bokeloh, Alireza Fathi, Bastian Leibe, Matthias Nießner; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9031-9040
We present 3D-MPA, a method for instance segmentation on 3D point clouds. Given an input point cloud, we propose an object-centric approach where each point votes for its object center. We sample object proposals from the predicted object centers. Then, we learn proposal features from grouped point features that voted for the same object center. A graph convolutional network introduces inter-proposal relations, providing higher-level feature learning in addition to the lower-level point features. Each proposal comprises a semantic label, a set of associated points over which we define a foreground-background mask, an objectness score and aggregation features. Previous works usually perform non-maximum-suppression (NMS) over proposals to obtain the final object detections or semantic instances. However, NMS can discard potentially correct predictions. Instead, our approach keeps all proposals and groups them together based on the learned aggregation features. We show that grouping proposals improves over NMS and outperforms previous state-of-the-art methods on the tasks of 3D object detection and semantic instance segmentation on the ScanNetV2 benchmark and the S3DIS dataset.
****

Domain Adaptive Image-to-Image Translation
Ying-Cong Chen, Xiaogang Xu, Jiaya Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5274-5283
Unpaired image-to-image translation (I2I) has achieved great success in various

applications. However, its generalization capacity is still an open question. In this paper, we show that existing I2I models do not generalize well for samples outside the training domain. The cause is twofold. First, an I2I model may not work well when testing samples are beyond its valid input domain. Second, results could be unreliable if the expected output is far from what the model is trained. To deal with these issues, we propose the Domain Adaptive Image-To-Image translation (DAI2I) framework that adapts an I2I model for out-of-domain samples. Our framework introduces two sub-modules -- one maps testing samples to the valid input domain of the I2I model, and the other transforms the output of I2I model to expected results. Extensive experiments manifest that our framework improves the capacity of existing I2I models, allowing them to handle samples that are distinctively different from their primary targets.

****

Video Playback Rate Perception for Self-Supervised Spatio-Temporal Representation Learning

Yuan Yao, Chang Liu, Dezhao Luo, Yu Zhou, Qixiang Ye; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6548-6557

In self-supervised spatio-temporal representation learning, the temporal resolution and long-short term characteristics are not yet fully explored, which limits representation capabilities of learned models. In this paper, we propose a novel self-supervised method, referred to as video Playback Rate Perception (PRP), to learn spatio-temporal representation in a simple-yet-effective way. PRP roots in a dilated sampling strategy, which produces self-supervision signals about video playback rates for representation model learning. PRP is implemented with a feature encoder, a classification module, and a reconstructing decoder, to achieve spatio-temporal semantic retention in a collaborative discrimination-generation manner. The discriminative perception model follows a feature encoder to prefer perceiving low temporal resolution and long-term representation by classifying fast-forward rates. The generative perception model acts as a feature decoder to focus on comprehending high temporal resolution and short-term representation by introducing a motion-attention mechanism. PRP is applied on typical video target tasks including action recognition and video retrieval. Experiments show that PRP outperforms state-of-the-art self-supervised models with significant margins. Code is available at github.com/yuanyao366/PRP.

****

Warping Residual Based Image Stitching for Large Parallax

Kyu-Yul Lee, Jae-Young Sim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8198-8206

Image stitching techniques align two images captured at different viewing positions onto a single wider image. When the captured 3D scene is not planar and the camera baseline is large, two images exhibit parallax where the relative positions of scene structures are quite different from each view. The existing image stitching methods often fail to work on the images with large parallax. In this paper, we propose an image stitching algorithm robust to large parallax based on the novel concept of warping residuals. We first estimate multiple homographies and find their inlier feature matches between two images. Then we evaluate warping residual for each feature match with respect to the multiple homographies. To alleviate the parallax artifacts, we partition input images into superpixels and warp each superpixel adaptively according to an optimal homography which is computed by minimizing the error of feature matches weighted by the warping residuals. Experimental results demonstrate that the proposed algorithm provides accurate stitching results for images with large parallax, and outperforms the existing methods qualitatively and quantitatively.

****

GLU-Net: Global-Local Universal Network for Dense Flow and Correspondences

Prune Truong, Martin Danelljan, Radu Timofte; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6258-6268

Establishing dense correspondences between a pair of images is an important and general problem, covering geometric matching, optical flow and semantic correspo

ndences. While these applications share fundamental challenges, such as large di
splacements, pixel-accuracy, and appearance changes, they are currently addresse
d with specialized network architectures, designed for only one particular task.
 This severely limits the generalization capabilities of such networks to new sc
enarios, where e.g. robustness to larger displacements or higher accuracy is req
uired. In this work, we propose a universal network architecture that is directl
y applicable to all the aforementioned dense correspondence problems. We achieve
 both high accuracy and robustness to large displacements by investigating the c
ombined use of global and local correlation layers. We further propose an adapti
ve resolution strategy, allowing our network to operate on virtually any input i
mage resolution. The proposed GLU-Net achieves state-of-the-art performance for
geometric and semantic matching as well as optical flow, when using the same net
work and weights. Code and trained models are available at https://github.com/Pr
uneTruong/GLU-Net.
****

SAINT: Spatially Aware Interpolation NeTwork for Medical Slice Synthesis
Cheng Peng, Wei-An Lin, Haofu Liao, Rama Chellappa, S. Kevin Zhou; Proceedin
gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
 2020, pp. 7750-7759
Deep learning-based single image super-resolution (SISR) methods face various ch
allenges when applied to 3D medical volumetric data (i.e., CT and MR images) due
 to the high memory cost and anisotropic resolution, which adversely affect thei
r performance. Furthermore, mainstream SISR methods are designed to work over sp
ecific upsampling factors, which makes them ineffective in clinical practice. In
 this paper, we introduce a Spatially Aware Interpolation NeTwork (SAINT) for me
dical slice synthesis to alleviate the memory constraint that volumetric data po
ses. Compared to other super-resolution methods, SAINT utilizes voxel spacing in
formation to provide desirable levels of details, and allows for the upsampling
factor to be determined on the fly. Our evaluations based on 853 CT scans from f
our datasets that contain liver, colon, hepatic vessels, and kidneys show that S
AINT consistently outperforms other SISR methods in terms of medical slice synth
esis quality, while using only a single model to deal with different upsampling
factors
****

StarGAN v2: Diverse Image Synthesis for Multiple Domains
Yunjey Choi, Youngjung Uh, Jaejun Yoo, Jung-Woo Ha; Proceedings of the IEEE/C
VF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8188-
8197
A good image-to-image translation model should learn a mapping between different
 visual domains while satisfying the following properties: 1) diversity of gener
ated images and 2) scalability over multiple domains. Existing methods address e
ither of the issues, having limited diversity or multiple models for all domains
. We propose StarGAN v2, a single framework that tackles both and shows signific
antly improved results over the baselines. Experiments on CelebA-HQ and a new an
imal faces dataset (AFHQ) validate our superiority in terms of visual quality, d
iversity, and scalability. To better assess image-to-image translation models, w
e release AFHQ, high-quality animal faces with large inter- and intra-domain dif
ferences. The code, pretrained models, and dataset are available at https://gith
ub.com/clovaai/stargan-v2.
****

Local Deep Implicit Functions for 3D Shape
Kyle Genova, Forrester Cole, Avneesh Sud, Aaron Sarna, Thomas Funkhouser; Pr
oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
(CVPR), 2020, pp. 4857-4866
The goal of this project is to learn a 3D shape representation that enables accu
rate surface reconstruction, compact storage, efficient computation, consistency
 for similar shapes, generalization across diverse shape categories, and inferen
ce from depth camera observations. Towards this end, we introduce Local Deep Imp
licit Functions (LDIF), a 3D shape representation that decomposes space into a s
tructured set of learned implicit functions. We provide networks that infer the

space decomposition and local deep implicit functions from a 3D mesh or posed de pth image. During experiments, we find that it provides 10.3 points higher surfa ce reconstruction accuracy (F-Score) than the state-of-the-art (OccNet), while r equiring fewer than 1% of the network parameters. Experiments on posed depth ima ge completion and generalization to unseen classes show 15.8 and 17.8 point impr ovements over the state-of-the-art, while producing a structured 3D representati on for each input with consistency across diverse shape collections.
****

Weakly-Supervised Domain Adaptation via GAN and Mesh Model for Estimating 3D Han d Poses Interacting Objects

Seungryul Baek, Kwang In Kim, Tae-Kyun Kim; Proceedings of the IEEE/CVF Confer ence on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6121-6131

Despite recent successes in hand pose estimation, there yet remain challenges on RGB-based 3D hand pose estimation (HPE) under hand-object interaction (HOI) sce narios where severe occlusions and cluttered backgrounds exhibit. Recent RGB HOI benchmarks have been collected either in real or synthetic domain, however, the size of datasets is far from enough to deal with diverse objects combined with hand poses, and 3D pose annotations of real samples are lacking, especially for occluded cases. In this work, we propose a novel end-to-end trainable pipeline t hat adapts the hand-object domain to the single hand-only domain, while learning for HPE. The domain adaption occurs in image space via 2D pixel-level guidance by Generative Adversarial Network (GAN) and 3D mesh guidance by mesh renderer (M R). Via the domain adaption in image space, not only 3D HPE accuracy is improved , but also HOI input images are translated to segmented and de-occluded hand-onl y images. The proposed method takes advantages of both the guidances: GAN accura tely aligns hands, while MR effectively fills in occluded pixels. The experiment s using Dexter-Object, Ego-Dexter and HO3D datasets show that our method signifi cantly outperforms state-of-the-arts trained by hand-only data and is comparable to those supervised by HOI data. Note our method is trained primarily by hand-o nly images with pose labels, and HOI images without pose labels.
****

Global Texture Enhancement for Fake Face Detection in the Wild

Zhengzhe Liu, Xiaojuan Qi, Philip H.S. Torr; Proceedings of the IEEE/CVF Confe rence on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8060-8069

Generative Adversarial Networks (GANs) can generate realistic fake face images t hat can easily fool human beings. On the contrary, a common Convolutional Neural Network(CNN) discriminator can achieve more than99.9%accuracyin discerning fake /real images. In this paper, we conduct an empirical study on fake/real faces, a nd have two important observations: firstly, the texture of fake faces is substa ntially different from real ones; secondly, global texture statistics are more r obust to image editing and transferable to fake faces from different GANs and da tasets. Motivated by the above observations, we propose a new architecture coine d as Gram-Net, which leverages global image texture representations for robust f ake image detection. Experimental results on several datasets demonstrate that o ur Gram-Netoutperforms existing approaches. Especially, our Gram-Netis more robu st to image editings, e.g. down-sampling, JPEGcompression, blur, and noise. More importantly, our Gram-Net generalizes significantly better in detecting fake fa ces from GAN models not seen in the training phase and can perform decently in d etecting fake natural images
****

C-Flow: Conditional Generative Flow Models for Images and 3D Point Clouds

Albert Pumarola, Stefan Popov, Francesc Moreno-Noguer, Vittorio Ferrari; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2020, pp. 7949-7958

Flow-based generative models have highly desirable properties like exact log-lik elihood evaluation and exact latent-variable inference, however they are still i n their infancy and have not received as much attention as alternative generativ e models. In this paper, we introduce C-Flow, a novel conditioning scheme that b rings normalizing flows to an entirely new scenario with great possibilities for multimodal data modeling. C-Flow is based on a parallel sequence of invertible

mappings in which a source flow guides the target flow at every step, enabling f ine-grained control over the generation process. We also devise a new strategy t o model unordered 3D point clouds that, in combination with the conditioning sch eme, makes it possible to address 3D reconstruction from a single image and its inverse problem of rendering an image given a point cloud. We demonstrate our co nditioning method to be very adaptable, being also applicable to image manipulat ion, style transfer and multi-modal image-to-image mapping in a diversity of dom ains, including RGB images, segmentation maps and edge masks.
****

Hyperbolic Image Embeddings
Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, Vi ctor Lempitsky; Proceedings of the IEEE/CVF Conference on Computer Vision and Pa ttern Recognition (CVPR), 2020, pp. 6418-6428
Computer vision tasks such as image classification, image retrieval, and few-sho t learning are currently dominated by Euclidean and spherical embeddings so that the final decisions about class belongings or the degree of similarity are made using linear hyperplanes, Euclidean distances, or spherical geodesic distances (cosine similarity). In this work, we demonstrate that in many practical scenari os, hyperbolic embeddings provide a better alternative.
****

Nested Scale-Editing for Conditional Image Synthesis
Lingzhi Zhang, Jiancong Wang, Yinshuang Xu, Jie Min, Tarmily Wen, James C. Gee, Jianbo Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5477-5487
We propose an image synthesis approach that provides stratified navigation in th e latent code space. With a tiny amount of partial or very low-resolution image, our approach can consistently out-perform state-of-the-art counterparts in term s of generating the closest sampled image to the ground truth. We achieve this t hrough scale-independent editing while expanding scale-specific diversity. Scale -independence is achieved with a nested scale disentanglement loss. Scale-specif ic diversity is created by incorporating a progressive diversification constrain t. We introduce semantic persistency across the scales by sharing common latent codes. Together they provide better control of the image synthesis process. We e valuate the effectiveness of our proposed approach through various tasks, includ ing image outpainting, image superresolution, and cross-domain image translation .
****

Joint Spatial-Temporal Optimization for Stereo 3D Object Tracking
Peiliang Li, Jieqi Shi, Shaojie Shen; Proceedings of the IEEE/CVF Conference o n Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6877-6886
Directly learning multiple 3D objects motion from sequential images is difficult , while the geometric bundle adjustment lacks the ability to localize the invisi ble object centroid. To benefit from both the powerful object understanding skil l from deep neural network meanwhile tackle precise geometry modeling for consis tent trajectory estimation, we propose a joint spatial-temporal optimization-bas ed stereo 3D object tracking method. From the network, we detect corresponding 2 D bounding boxes on adjacent images and regress an initial 3D bounding box. Dens e object cues (local depth and local coordinates) that associating to the object centroid are then predicted using a region-based network. Considering both the instant localization accuracy and motion consistency, our optimization models th e relations between the object centroid and observed cues into a joint spatial-t emporal error function. All historic cues will be summarized to contribute to th e current estimation by a per-frame marginalization strategy without repeated co mputation. Quantitative evaluation on the KITTI tracking dataset shows our appro ach outperforms previous image-based 3D tracking methods by significant margins. We also report extensive results on multiple categories and larger datasets (KI TTI raw and Argoverse Tracking) for future benchmarking.
****

Reusing Discriminators for Encoding: Towards Unsupervised Image-to-Image Transla tion

Runfa Chen, Wenbing Huang, Binghui Huang, Fuchun Sun, Bin Fang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8168-8177

Unsupervised image-to-image translation is a central task in computer vision. Current translation frameworks will abandon the discriminator once the training process is completed. This paper contends a novel role of the discriminator by reusing it for encoding the images of the target domain. The proposed architecture, termed as NICE-GAN, exhibits two advantageous patterns over previous approaches: First, it is more compact since no independent encoding component is required; Second, this plug-in encoder is directly trained by the adversary loss, making it more informative and trained more effectively if a multi-scale discriminator is applied. The main issue in NICE-GAN is the coupling of translation with discrimination along the encoder, which could incur training inconsistency when we play the min-max game via GAN. To tackle this issue, we develop a decoupled training strategy by which the encoder is only trained when maximizing the adversary loss while keeping frozen otherwise. Extensive experiments on four popular benchmarks demonstrate the superior performance of NICE-GAN over state-of-the-art methods in terms of FID, KID, and also human preference. Comprehensive ablation studies are also carried out to isolate the validity of each proposed component. Our codes are available at https://github.com/alpc91/NICE-GAN-pytorch.

****

Learning Representations by Predicting Bags of Visual Words

Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Perez, Matthieu Cord; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6928-6938

Self-supervised representation learning targets to learn convnet-based image representations from unlabeled data. Inspired by the success of NLP methods in this area, in this work we propose a self-supervised approach based on spatially dense image descriptions that encode discrete visual concepts, here called visual words. To build such discrete representations, we quantize the feature maps of a first pre-trained self-supervised convnet, over a k-means based vocabulary. Then, as a self-supervised task, we train another convnet to predict the histogram of visual words of an image (i.e., its Bag-of-Words representation) given as input a perturbed version of that image. The proposed task forces the convnet to learn perturbation-invariant and context-aware image features, useful for downstream image understanding tasks. We extensively evaluate our method and demonstrate very strong empirical results, e.g., our pre-trained self-supervised representations transfer better on detection task and similarly on classification over classes "unseen" during pre-training, when compared to the supervised case. This also shows that the process of image discretization into visual words can provide the basis for very powerful self-supervised approaches in the image domain, thus allowing further connections to be made to related methods from the NLP domain that have been extremely successful so far.

****

Global-Local Bidirectional Reasoning for Unsupervised Representation Learning of 3D Point Clouds

Yongming Rao, Jiwen Lu, Jie Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5376-5385

Local and global patterns of an object are closely related. Although each part of an object is incomplete, the underlying attributes about the object are shared among all parts, which makes reasoning the whole object from a single part possible. We hypothesize that a powerful representation of a 3D object should model the attributes that are shared between parts and the whole object, and distinguishable from other objects. Based on this hypothesis, we propose to learn point cloud representation by bidirectional reasoning between the local structures at different abstraction hierarchies and the global shape without human supervision. Experimental results on various benchmark datasets demonstrate the unsupervisedly learned representation is even better than supervised representation in discriminative power, generalization ability, and robustness. We show that unsupervisedly trained point cloud models can outperform their supervised counterparts on

downstream classification tasks. Most notably, by simply increasing the channel width of an SSG PointNet++, our unsupervised model surpasses the state-of-the-art supervised methods on both synthetic and real-world 3D object classification datasets. We expect our observations to offer a new perspective on learning better representation from data structures instead of human annotations for point cloud understanding.

****

Knowledge As Priors: Cross-Modal Knowledge Generalization for Datasets Without Superior Knowledge

Long Zhao, Xi Peng, Yuxiao Chen, Mubbasir Kapadia, Dimitris N. Metaxas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6528-6537

Cross-modal knowledge distillation deals with transferring knowledge from a model trained with superior modalities (Teacher) to another model trained with weak modalities (Student). Existing approaches require paired training examples exist in both modalities. However, accessing the data from superior modalities may not always be feasible. For example, in the case of 3D hand pose estimation, depth maps, point clouds, or stereo images usually capture better hand structures than RGB images, but most of them are expensive to be collected. In this paper, we propose a novel scheme to train the Student in a Target dataset where the Teacher is unavailable. Our key idea is to generalize the distilled cross-modal knowledge learned from a Source dataset, which contains paired examples from both modalities, to the Target dataset by modeling knowledge as priors on parameters of the Student. We name our method "Cross-Modal Knowledge Generalization" and demonstrate that our scheme results in competitive performance for 3D hand pose estimation on standard benchmark datasets.

****

Large Scale Video Representation Learning via Relational Graph Clustering

Hyodong Lee, Joonseok Lee, Joe Yue-Hei Ng, Paul Natsev; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6807-6816

Representation learning is widely applied for various tasks on multimedia data, e.g., retrieval and search. One approach for learning useful representation is by utilizing the relationships or similarities between examples. In this work, we explore two promising scalable representation learning approaches on video domain. With hierarchical graph clusters built upon video-to-video similarities, we propose: 1) smart negative sampling strategy that significantly boosts training efficiency with triplet loss, and 2) a pseudo-classification approach using the clusters as pseudo-labels. The embeddings trained with the proposed methods are competitive on multiple video understanding tasks, including related video retrieval and video annotation. Both of these proposed methods are highly scalable, as verified by experiments on large-scale datasets.

****

ASLFeat: Learning Local Features of Accurate Shape and Localization

Zixin Luo, Lei Zhou, Xuyang Bai, Hongkai Chen, Jiahui Zhang, Yao Yao, Shiwei Li, Tian Fang, Long Quan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6589-6598

This work focuses on mitigating two limitations in the joint learning of local feature detectors and descriptors. First, the ability to estimate the local shape (scale, orientation, etc.) of feature points is often neglected during dense feature extraction, while the shape-awareness is crucial to acquire stronger geometric invariance. Second, the localization accuracy of detected keypoints is not sufficient to reliably recover camera geometry, which has become the bottleneck in tasks such as 3D reconstruction. In this paper, we present ASLFeat, with three light-weight yet effective modifications to mitigate above issues. First, we resort to deformable convolutional networks to densely estimate and apply local transformation. Second, we take advantage of the inherent feature hierarchy to restore spatial resolution and low-level details for accurate keypoint localization. Finally, we use a peakiness measurement to relate feature responses and derive more indicative detection scores. The effect of each modification is thoroughl

y studied, and the evaluation is extensively conducted across a variety of practical scenarios. State-of-the-art results are reported that demonstrate the superiority of our methods.
****

Video Super-Resolution With Temporal Group Attention
Takashi Isobe, Songjiang Li, Xu Jia, Shanxin Yuan, Gregory Slabaugh, Chunjing Xu, Ya-Li Li, Shengjin Wang, Qi Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8008-8017
Video super-resolution, which aims at producing a high-resolution video from its corresponding low-resolution version, has recently drawn increasing attention. In this work, we propose a novel method that can effectively incorporate temporal information in a hierarchical way. The input sequence is divided into several groups, with each one corresponding to a kind of frame rate. These groups provide complementary information to recover missing details in the reference frame, which is further integrated with an attention module and a deep intra-group fusion module. In addition, a fast spatial alignment is proposed to handle videos with large motion. Extensive results demonstrate the capability of the proposed model in handling videos with various motion. It achieves favorable performance against state-of-the-art methods on several benchmark datasets.
****

TailorNet: Predicting Clothing in 3D as a Function of Human Pose, Shape and Garment Style
Chaitanya Patel, Zhouyingcheng Liao, Gerard Pons-Moll; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7365-7375
In this paper, we present TailorNet, a neural model which predicts clothing deformation in 3D as a function of three factors: pose, shape and style (garment geometry), while retaining wrinkle detail. This goes beyond prior models, which are either specific to one style and shape, or generalize to different shapes producing smooth results, despite being style specific. Our hypothesis is that (even non-linear) combinations of examples smoothes out high frequency components such as fine-wrinkles, which makes learning the three factors jointly hard. At the heart of our technique is a decomposition of deformation into a high frequency and a low frequency component. While the low-frequency component is predicted from pose, shape and style parameters with an MLP, the high-frequency component is predicted with a mixture of shape-style specific pose models. The weights of the mixture are computed with a narrow bandwidth kernel to guarantee that only predictions with similar high-frequency patterns are combined. The style variation is obtained by computing, in a canonical pose, a subspace of deformation, which satisfies physical constraints such as inter-penetration, and draping on the body. TailorNet delivers 3D garments which retain the wrinkles from the physics based simulations (PBS) it is learned from, while running more than 1000 times faster. In contrast to classical PBS, TailorNet is easy to use and fully differentiable, which is crucial for computer vision and learning algorithms. Several experiments demonstrate TailorNet produces more realistic results than prior work, and even generates temporally coherent deformations on sequences of the AMASS dataset, despite being trained on static poses from a different dataset. To stimulate further research in this direction, we will make a dataset consisting of 55800 frames, as well as our model publicly available at https://virtualhumans.mpi-inf.mpg.de/tailornet/.
****

CurricularFace: Adaptive Curriculum Learning Loss for Deep Face Recognition
Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, Feiyue Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5901-5910
As an emerging topic in face recognition, designing margin-based loss functions can increase the feature margin between different classes for enhanced discriminability. More recently, the idea of mining-based strategies is adopted to emphasize the misclassified samples, achieving promising results. However, during the entire training process, the prior methods either do not explicitly emphasize th

e sample based on its importance that renders the hard samples not fully exploit ed; or explicitly emphasize the effects of semi-hard/hard samples even at the ea rly training stage that may lead to convergence issue. In this work, we propose a novel Adaptive Curriculum Learning loss (CurricularFace) that embeds the idea of curriculum learning into the loss function to achieve a novel training strate gy for deep face recognition, which mainly addresses easy samples in the early t raining stage and hard ones in the later stage. Specifically, our CurricularFace adaptively adjusts the relative importance of easy and hard samples during diff erent training stages. In each stage, different samples are assigned with differ ent importance according to their corresponding difficultness. Extensive experim ental results on popular benchmarks demonstrate the superiority of our Curricula rFace over the state-of-the-art competitors.

****

On the Detection of Digital Face Manipulation

Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, Anil K. Jain; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20 20, pp. 5781-5790

Detecting manipulated facial images and videos is an increasingly important topi c in digital media forensics. As advanced face synthesis and manipulation method s are made available, new types of fake face representations are being created w hich have raised significant concerns for their use in social media. Hence, it i s crucial to detect manipulated face images and localize manipulated regions. In stead of simply using multi-task learning to simultaneously detect manipulated i mages and predict the manipulated mask (regions), we propose to utilize an atten tion mechanism to process and improve the feature maps for the classification ta sk. The learned attention maps highlight the informative regions to further impr ove the binary classification (genuine face v. fake face), and also visualize th e manipulated regions. To enable our study of manipulated face detection and loc alization, we collect a large-scale database that contains numerous types of fac ial forgeries. With this dataset, we perform a thorough analysis of data-driven fake face detection. We show that the use of an attention mechanism improves fac ial forgery detection and manipulated region localization.

****

Sketch-BERT: Learning Sketch Bidirectional Encoder Representation From Transform ers by Self-Supervised Learning of Sketch Gestalt

Hangyu Lin, Yanwei Fu, Xiangyang Xue, Yu-Gang Jiang; Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6758 -6767

Previous researches of sketches often considered sketches in pixel format and le veraged CNN based models in the sketch understanding. Fundamentally, a sketch is stored as a sequence of data points, a vector format representation, rather tha n the photo-realistic image of pixels. SketchRNN studied a generative neural rep resentation for sketches of vector format by Long Short Term Memory networks (LS TM). Unfortunately, the representation learned by SketchRNN is primarily for the generation tasks, rather than the other tasks of recognition and retrieval of s ketches. To this end and inspired by the recent BERT model, we present a model o f learning Sketch Bidirectional Encoder Representation from Transformer (Sketch- BERT). We generalize BERT to sketch domain, with the novel proposed components a nd pre-training algorithms, including the newly designed sketch embedding networ ks, and the self-supervised learning of sketch gestalt. Particularly, towards th e pre-training task, we present a novel Sketch Gestalt Model (SGM) to help train the Sketch-BERT. Experimentally, we show that the learned representation of Ske tch-BERT can help and improve the performance of the downstream tasks of sketch recognition, sketch retrieval, and sketch gestalt.

****

Decoupled Representation Learning for Skeleton-Based Gesture Recognition

Jianbo Liu, Yongcheng Liu, Ying Wang, Veronique Prinet, Shiming Xiang, Chun hong Pan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5751-5760

Skeleton-based gesture recognition is very challenging, as the high-level inform

ation in gesture is expressed by a sequence of complexly composite motions. Previous works often learn all the motions with a single model. In this paper, we propose to decouple the gesture into hand posture variations and hand movements, which are then modeled separately. For the former, the skeleton sequence is embedded into a 3D hand posture evolution volume (HPEV) to represent fine-grained posture variations. For the latter, the shifts of hand center and fingertips are arranged as a 2D hand movement map (HMM) to capture holistic movements. To learn from the two inhomogeneous representations for gesture recognition, we propose an end-to-end two-stream network. The HPEV stream integrates both spatial layout and temporal evolution information of hand postures by a dedicated 3D CNN, while the HMM stream develops an efficient 2D CNN to extract hand movement features. Eventually, the predictions of the two streams are aggregated with high efficiency. Extensive experiments on SHREC'17 Track, DHG-14/28 and FPHA datasets demonstrate that our method is competitive with the state-of-the-art.

****

Analyzing and Improving the Image Quality of StyleGAN

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, Timo Aila; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8110-8119

The style-based GAN architecture (StyleGAN) yields state-of-the-art results in data-driven unconditional generative image modeling. We expose and analyze several of its characteristic artifacts, and propose changes in both model architecture and training methods to address them. In particular, we redesign the generator normalization, revisit progressive growing, and regularize the generator to encourage good conditioning in the mapping from latent codes to images. In addition to improving image quality, this path length regularizer yields the additional benefit that the generator becomes significantly easier to invert. This makes it possible to reliably attribute a generated image to a particular network. We furthermore visualize how well the generator utilizes its output resolution, and identify a capacity problem, motivating us to train larger models for additional quality improvements. Overall, our improved model redefines the state of the art in unconditional image modeling, both in terms of existing distribution quality metrics as well as perceived image quality.

****

Learning to Dress 3D People in Generative Clothing

Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, Michael J. Black; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6469-6478

Three-dimensional human body models are widely used in the analysis of human pose and motion. Existing models, however, are learned from minimally-clothed 3D scans and thus do not generalize to the complexity of dressed people in common images and videos. Additionally, current models lack the expressive power needed to represent the complex non-linear geometry of pose-dependent clothing shapes. To address this, we learn a generative 3D mesh model of clothed people from 3D scans with varying pose and clothing. Specifically, we train a conditional Mesh-VAE-GAN to learn the clothing deformation from the SMPL body model, making clothing an additional term in SMPL. Our model is conditioned on both pose and clothing type, giving the ability to draw samples of clothing to dress different body shapes in a variety of styles and poses. To preserve wrinkle detail, our Mesh-VAE-GAN extends patchwise discriminators to 3D meshes. Our model, named CAPE, represents global shape and fine local structure, effectively extending the SMPL body model to clothing. To our knowledge, this is the first generative model that directly dresses 3D human body meshes and generalizes to different poses. The model, code and data are available for research purposes at https://cape.is.tue.mpg.de.

****

Cross-Modal Pattern-Propagation for RGB-T Tracking

Chaoqun Wang, Chunyan Xu, Zhen Cui, Ling Zhou, Tong Zhang, Xiaoya Zhang, Jian Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7064-7073

Motivated by our observations on RGB-T data that pattern correlations are high-frequently recurred across modalities also along sequence frames, in this paper, we propose a cross-modal pattern-propagation (CMPP) tracking framework to diffuse instance patterns across RGB-T data on spatial domain as well as temporal domain. To bridge RGB-T modalities, the cross-modal correlations on intra-modal paired pattern-affinities are derived to reveal those latent cues between heterogeneous modalities. Through the correlations, the useful patterns may be mutually propagated between RGB-T modalities so as to fulfill inter-modal pattern-propagation. Further, considering the temporal continuity of sequence frames, we adopt the spirit of pattern propagation to dynamic temporal domain, in which long-term historical contexts are adaptively correlated and propagated into the current frame for more effective information inheritance. Extensive experiments demonstrate that the effectiveness of our proposed CMPP, and the new state-of-the-art results are achieved with the significant improvements on two RGB-T object tracking benchmarks.

****

Channel Attention Based Iterative Residual Learning for Depth Map Super-Resolution

Xibin Song, Yuchao Dai, Dingfu Zhou, Liu Liu, Wei Li, Hongdong Li, Ruigang Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5631-5640

Despite the remarkable progresses made in deep learning based depth map super-resolution (DSR), how to tackle real-world degradation in low-resolution (LR) depth maps remains a major challenge. Existing DSR model is generally trained and tested on synthetic dataset, which is very different from what would get from a real depth sensor. In this paper, we argue that DSR models trained under this setting are restrictive and not effective in dealing with realworld DSR tasks. We make two contributions in tackling real-world degradation of different depth sensors. First, we propose to classify the generation of LR depth maps into two types : non-linear downsampling with noise and interval downsampling, for which DSR models are learned correspondingly. Second, we propose a new framework for real-world DSR, which consists of four modules : 1) An iterative residual learning module with deep supervision to learn effective high-frequency components of depth maps in a coarse-to-fine manner; 2) A channel attention strategy to enhance channels with abundant high-frequency components; 3) A multi-stage fusion module to effectively reexploit the results in the coarse-to-fine process; and 4) A depth refinement module to improve the depth map by TGV regularization and input loss. Extensive experiments on benchmarking datasets demonstrate the superiority of our method over current state-of-the-art DSR methods.

****

Averaging Essential and Fundamental Matrices in Collinear Camera Settings

Amnon Geifman, Yoni Kasten, Meirav Galun, Ronen Basri; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6021-6030

Global methods to Structure from Motion have gained popularity in recent years. A significant drawback of global methods is their sensitivity to collinear camera settings. In this paper, we introduce an analysis and algorithms for averaging bifocal tensors (essential or fundamental matrices) when either subsets or all of the camera centers are collinear. We provide a complete spectral characterization of bifocal tensors in collinear scenarios and further propose two averaging algorithms. The first algorithm uses rank constrained minimization to recover camera matrices in fully collinear settings. The second algorithm enriches the set of possibly mixed collinear and non-collinear cameras with additional, "virtual cameras," which are placed in general position, enabling the application of existing averaging methods to the enriched set of bifocal tensors. Our algorithms are shown to achieve state of the art results on various benchmarks that include autonomous car datasets and unordered image collections in both calibrated and unclibrated settings.

****

Deep Spatial Gradient and Temporal Depth Learning for Face Anti-Spoofing

Zezheng Wang, Zitong Yu, Chenxu Zhao, Xiangyu Zhu, Yunxiao Qin, Qiusheng Zhou, Feng Zhou, Zhen Lei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5042-5051

Face anti-spoofing is critical to the security of face recognition systems. Depth supervised learning has been proven as one of the most effective methods for face anti-spoofing. Despite the great success, most previous works still formulate the problem as a single-frame multi-task one by simply augmenting the loss with depth, while neglecting the detailed fine-grained information and the interplay between facial depths and moving patterns. In contrast, we design a new approach to detect presentation attacks from multiple frames based on two insights: 1) detailed discriminative clues (e.g., spatial gradient magnitude) between living and spoofing face may be discarded through stacked vanilla convolutions, and 2) the dynamics of 3D moving faces provide important clues in detecting the spoofing faces. The proposed method is able to capture discriminative details via Residual Spatial Gradient Block (RSGB) and encode spatio-temporal information from Spatio-Temporal Propagation Module (STPM) efficiently. Moreover, a novel Contrastive Depth Loss is presented for more accurate depth supervision. To assess the efficacy of our method, we also collect a Double-modal Anti-spoofing Dataset (DMAD) which provides actual depth for each sample. The experiments demonstrate that the proposed approach achieves state-of-the-art results on five benchmark datasets including OULU-NPU, SiW, CASIA-MFSD, Replay-Attack, and the new DMAD. Codes will be available at https://github.com/clks-wzz/FAS-SGTD.

****

Instance-Aware Image Colorization

Jheng-Wei Su, Hung-Kuo Chu, Jia-Bin Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7968-7977

Image colorization is inherently an ill-posed problem with multi-modal uncertainty. Previous methods leverage the deep neural network to map input grayscale images to plausible color outputs directly. Although these learning-based methods have shown impressive performance, they usually fail on the input images that contain multiple objects. The leading cause is that existing models perform learning and colorization on the entire image. In the absence of a clear figure-ground separation, these models cannot effectively locate and learn meaningful object-level semantics. In this paper, we propose a method for achieving instance-aware colorization. Our network architecture leverages an off-the-shelf object detector to obtain cropped object images and uses an instance colorization network to extract object-level features. We use a similar network to extract the full-image features and apply a fusion module to full object-level and image-level features to predict the final colors. Both colorization networks and fusion modules are learned from a large-scale dataset. Experimental results show that our work outperforms existing methods on different quality metrics and achieves state-of-the-art performance on image colorization.

****

ReDA:Reinforced Differentiable Attribute for 3D Face Reconstruction

Wenbin Zhu, HsiangTao Wu, Zeyu Chen, Noranart Vesdapunt, Baoyuan Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4958-4967

The key challenge for 3D face shape reconstruction is to build the correct dense face correspondence between the deformable mesh and the single input image. Given the ill-posed nature, previous works heavily rely on prior knowledge (such as 3DMM [2]) to reduce depth ambiguity. Although impressive result has been made recently [42, 14, 8], there is still a large room to improve the correspondence so that projected face shape better aligns with the silhouette of each face region (i.e, eye, mouth, nose, cheek, etc.) on the image. To further reduce the ambiguities, we present a novel framework called "Reinforced Differentiable Attributes" ("ReDA") which is more general and effective than previous Differentiable Rendering ("DR"). Specifically, we first extend from color to more broad attributes, including the depth and the face parsing mask. Secondly, unlike the previous Z-buffer rendering, we make the rendering to be more differentiable through a set of convolution operations with multi-scale kernel sizes. In the meanwhile, to m

ake "ReDA" to be more successful for 3D face recon-struction, we further introduce a new free-form deformation layer that sits on top of 3DMM to enjoy both the prior knowledge and out-of-space modeling. Both techniques can be easily integrated into existing 3D face reconstruction pipeline. Extensive experiments on both RGB and RGB-D datasets show that our approach outperforms prior arts.
****

Towards Global Explanations of Convolutional Neural Networks With Concept Attribution
Weibin Wu, Yuxin Su, Xixian Chen, Shenglin Zhao, Irwin King, Michael R. Lyu, Yu-Wing Tai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8652-8661
With the growing prevalence of convolutional neural networks (CNNs), there is an urgent demand to explain their behaviors. Global explanations contribute to understanding model predictions on a whole category of samples, and thus have attracted increasing interest recently. However, existing methods overwhelmingly conduct separate input attribution or rely on local approximations of models, making them fail to offer faithful global explanations of CNNs. To overcome such drawbacks, we propose a novel two-stage framework, Attacking for Interpretability (AfI), which explains model decisions in terms of the importance of user-defined concepts. AfI first conducts a feature occlusion analysis, which resembles a process of attacking models to derive the category-wide importance of different features. We then map the feature importance to concept importance through ad-hoc semantic tasks. Experimental results confirm the effectiveness of AfI and its superiority in providing more accurate estimations of concept importance than existing proposals.
****

Cross-Domain Face Presentation Attack Detection via Multi-Domain Disentangled Representation Learning
Guoqing Wang, Hu Han, Shiguang Shan, Xilin Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6678-6687
Face presentation attack detection (PAD) has been an urgent problem to be solved in the face recognition systems. Conventional approaches usually assume the testing and training are within the same domain; as a result, they may not generalize well into unseen scenarios because the representations learned for PAD may overfit to the subjects in the training set. In light of this, we propose an efficient disentangled representation learning for cross-domain face PAD. Our approach consists of disentangled representation learning (DR-Net) and multi-domain learning (MD-Net). DR-Net learns a pair of encoders via generative models that can disentangle PAD informative features from subject discriminative features. The disentangled features from different domains are fed to MD-Net which learns domain-independent features for the final cross-domain face PAD task. Extensive experiments on several public datasets validate the effectiveness of the proposed approach for cross-domain PAD.
****

Time Flies: Animating a Still Image With Time-Lapse Video As Reference
Chia-Chi Cheng, Hung-Yu Chen, Wei-Chen Chiu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5641-5650
Time-lapse videos usually perform eye-catching appearances but are often hard to create. In this paper, we propose a self-supervised end-to-end model to generate the time-lapse video from a single image and a reference video. Our key idea is to extract both the style and the features of temporal variation from the reference video, and transfer them onto the input image. To ensure both the temporal consistency and realness of our resultant videos, we introduce several novel designs in our architecture, including classwise NoiseAdaIN, flow loss, and the video discriminator. In comparison to the baselines of state-of-the-art style transfer approaches, our proposed method is not only efficient in computation but also able to create more realistic and temporally smooth time-lapse video of a still image, with its temporal variation consistent to the reference.
****

PandaNet: Anchor-Based Single-Shot Multi-Person 3D Pose Estimation
Abdallah Benzine, Florian Chabot, Bertrand Luvison, Quoc Cuong Pham, Catherine Achard; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6856-6865
Recently, several deep learning models have been proposed for 3D human pose estimation. Nevertheless, most of these approaches only focus on the single-person case or estimate 3D pose of a few people at high resolution. Furthermore, many applications such as autonomous driving or crowd analysis require pose estimation of a large number of people possibly at low-resolution. In this work, we present PandaNet (Pose estimAtioN and Dectection Anchor-based Network), a new single-shot, anchor-based and multi-person 3D pose estimation approach. The proposed model performs bounding box detection and, for each detected person, 2D and 3D pose regression into a single forward pass. It does not need any post-processing to regroup joints since the network predicts a full 3D pose for each bounding box and allows the pose estimation of a possibly large number of people at low resolution. To manage people overlapping, we introduce a Pose-Aware Anchor Selection strategy. Moreover, as imbalance exists between different people sizes in the image, and joints coordinates have different uncertainties depending on these sizes, we propose a method to automatically optimize weights associated to different people scales and joints for efficient training. PandaNet surpasses previous single-shot methods on several challenging datasets: a multi-person urban virtual but very realistic dataset (JTA Dataset), and two real world 3D multi-person datasets (CMU Panoptic and MuPoTS-3D).
****
Modeling the Background for Incremental Learning in Semantic Segmentation
Fabio Cermelli, Massimiliano Mancini, Samuel Rota Bulo, Elisa Ricci, Barbara Caputo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9233-9242
Despite their effectiveness in a wide range of tasks, deep architectures suffer from some important limitations. In particular, they are vulnerable to catastrophic forgetting, i.e. they perform poorly when they are required to update their model as new classes are available but the original training set is not retained. This paper addresses this problem in the context of semantic segmentation. Current strategies fail on this task because they do not consider a peculiar aspect of semantic segmentation: since each training step provides annotation only for a subset of all possible classes, pixels of the background class (i.e. pixels that do not belong to any other classes) exhibit a semantic distribution shift. In this work we revisit classical incremental learning methods, proposing a new distillation-based framework which explicitly accounts for this shift. Furthermore, we introduce a novel strategy to initialize classifier's parameters, thus preventing biased predictions toward the background class. We demonstrate the effectiveness of our approach with an extensive evaluation on the Pascal-VOC 2012 and ADE20K datasets, significantly outperforming state of the art incremental learning methods.
****
F-BRS: Rethinking Backpropagating Refinement for Interactive Segmentation
Konstantin Sofiiuk, Ilia Petrov, Olga Barinova, Anton Konushin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8623-8632
Deep neural networks have become a mainstream approach to interactive segmentation. As we show in our experiments, while for some images a trained network provides accurate segmentation result with just a few clicks, for some unknown objects it cannot achieve satisfactory result even with a large amount of user input. Recently proposed backpropagating refinement scheme (BRS) introduces an optimization problem for interactive segmentation that results in significantly better performance for the hard cases. At the same time, BRS requires running forward and backward pass through a deep network several times that leads to significantly increased computational budget per click compared to other methods. We propose f-BRS (feature backpropagating refinement scheme) that solves an optimization problem with respect to auxiliary variables instead of the network inputs, and req

uires running forward and backward passes just for a small part of a network. Experiments on GrabCut, Berkeley, DAVIS and SBD datasets set new state-of-the-art at an order of magnitude lower time per click compared to original BRS. The code and trained models are available at https://github.com/saic-vul/fbrs_interactive_segmentation.
****

Steering Self-Supervised Feature Learning Beyond Local Pixel Statistics
Simon Jenni, Hailin Jin, Paolo Favaro; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6408-6417
We introduce a novel principle for self-supervised feature learning based on the discrimination of specific transformations of an image. We argue that the generalization capability of learned features depends on what image neighborhood size is sufficient to discriminate different image transformations: The larger the required neighborhood size and the more global the image statistics that the feature can describe. An accurate description of global image statistics allows to better represent the shape and configuration of objects and their context, which ultimately generalizes better to new tasks such as object classification and detection. This suggests a criterion to choose and design image transformations. Based on this criterion, we introduce a novel image transformation that we call limited context inpainting (LCI). This transformation inpaints an image patch conditioned only on a small rectangular pixel boundary (the limited context). Because of the limited boundary information, the inpainter can learn to match local pixel statistics, but is unlikely to match the global statistics of the image. We claim that the same principle can be used to justify the performance of transformations such as image rotations and warping. Indeed, we demonstrate experimentally that learning to discriminate transformations such as LCI, image warping and rotations, yields features with state of the art generalization capabilities on several datasets such as Pascal VOC, STL-10, CelebA, and ImageNet. Remarkably, our trained features achieve a performance on Places on par with features trained through supervised learning with ImageNet labels.
****

Weakly-Supervised Mesh-Convolutional Hand Reconstruction in the Wild
Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M. Bronstein, Stefanos Zafeiriou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4990-5000
We introduce a simple and effective network architecture for monocular 3D hand pose estimation consisting of an image encoder followed by a mesh convolutional decoder that is trained through a direct 3D hand mesh reconstruction loss. We train our network by gathering a large-scale dataset of hand action in YouTube videos and use it as a source of weak supervision. Our weakly-supervised mesh convolutions-based system largely outperforms state-of-the-art methods, even halving the errors on the in the wild benchmark. The dataset and additional resources are available at https://arielai.com/mesh_hands.
****

Reinforced Feature Points: Optimizing Feature Detection and Description for a High-Level Task
Aritra Bhowmik, Stefan Gumhold, Carsten Rother, Eric Brachmann; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4948-4957
We address a core problem of computer vision: Detection and description of 2D feature points for image matching. For a long time, hand-crafted designs, like the seminal SIFT algorithm, were unsurpassed in accuracy and efficiency. Recently, learned feature detectors emerged that implement detection and description using neural networks. Training these networks usually resorts to optimizing low-level matching scores, often pre-defining sets of image patches which should or should not match, or which should or should not contain key points. Unfortunately, increased accuracy for these low-level matching scores does not necessarily translate to better performance in high-level vision tasks. We propose a new training methodology which embeds the feature detector in a complete vision pipeline, and where the learnable parameters are trained in an end-to-end fashion. We overco

me the discrete nature of key point selection and descriptor matching using prin ciples from reinforcement learning. As an example, we address the task of relati ve pose estimation between a pair of images. We demonstrate that the accuracy of a state-of-the-art learning-based feature detector can be increased when traine d for the task it is supposed to solve at test time. Our training methodology po ses little restrictions on the task to learn, and works for any architecture whi ch predicts key point heat maps, and descriptors for key point locations.
****

ProAlignNet: Unsupervised Learning for Progressively Aligning Noisy Contours
VSR Veeravasarapu, Abhishek Goel, Deepak Mittal, Maneesh Singh; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202 0, pp. 9671-9679
Contour shape alignment is a fundamental but challenging problem in computer vis ion, especially when the observations are partial, noisy, and largely misaligned . Recent ConvNet-based architectures that were proposed to align image structure s tend to fail with contour representation of shapes, mostly due to the use of p roximity-insensitive pixel-wise similarity measures as loss functions in their t raining processes. This work presents a novel ConvNet, "ProAlignNet," that accou nts for large scale misalignments and complex transformations between the contou r shapes. It infers the warp parameters in a multi-scale fashion with progressiv ely increasing complex transformations over increasing scales. It learns --witho ut supervision-- to align contours, agnostic to noise and missing parts, by trai ning with a novel loss function which is derived an upperbound of a proximity-se nsitive and local shape-dependent similarity metric that uses classical Morpholo gical Chamfer Distance Transform. We evaluate the reliability of these proposals on a simulated MNIST noisy contours dataset via some basic sanity check experim ents. Next, we demonstrate the effectiveness of the proposed models in two real- world applications of (i) aligning geo-parcel data to aerial image maps and (ii) refining coarsely annotated segmentation labels. In both applications, the prop osed models consistently perform superior to state-of-the-art methods.
****

Attentive Normalization for Conditional Image Generation
Yi Wang, Ying-Cong Chen, Xiangyu Zhang, Jian Sun, Jiaya Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5094-5103
Traditional convolution-based generative adversarial networks synthesize images based on hierarchical local operations, where long-range dependency relation is implicitly modeled with a Markov chain. It is still not sufficient for categorie s with complicated structures. In this paper, we characterize long-range depende nce with attentive normalization (AN), which is an extension to traditional inst ance normalization. Specifically, the input feature map is softly divided into s everal regions based on its internal semantic similarity, which are respectively normalized. It enhances consistency between distant regions with semantic corre spondence. Compared with self-attention GAN, our attentive normalization does no t need to measure the correlation of all locations, and thus can be directly app lied to large-size feature maps without much computational burden. Extensive exp eriments on class-conditional image generation and semantic inpainting verify th e efficacy of our proposed module.
****

Learning by Analogy: Reliable Supervision From Transformations for Unsupervised Optical Flow Estimation
Liang Liu, Jiangning Zhang, Ruifei He, Yong Liu, Yabiao Wang, Ying Tai, Do nghao Luo, Chengjie Wang, Jilin Li, Feiyue Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6489-64 98
Unsupervised learning of optical flow, which leverages the supervision from view synthesis, has emerged as a promising alternative to supervised methods. Howeve r, the objective of unsupervised learning is likely to be unreliable in challeng ing scenes. In this work, we present a framework to use more reliable supervisio n from transformations. It simply twists the general unsupervised learning pipel

ine by running another forward pass with transformed data from augmentation, alo
ng with using transformed predictions of original data as the self-supervision s
ignal. Besides, we further introduce a lightweight network with multiple frames
by a highly-shared flow decoder. Our method consistently gets a leap of performa
nce on several benchmarks with the best accuracy among deep unsupervised methods
. Also, our method achieves competitive results to recent fully supervised metho
ds while with much fewer parameters.
****
Towards Better Generalization: Joint Depth-Pose Learning Without PoseNet
Wang Zhao, Shaohui Liu, Yezhi Shu, Yong-Jin Liu; Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9151-916
1
In this work, we tackle the essential problem of scale inconsistency for self su
pervised joint depth-pose learning. Most existing methods assume that a consiste
nt scale of depth and pose can be learned across all input samples, which makes
the learning problem harder, resulting in degraded performance and limited gener
alization in indoor environments and long-sequence visual odometry application.
To address this issue, we propose a novel system that explicitly disentangles sc
ale from the network estimation. Instead of relying on PoseNet architecture, our
 method recovers relative pose by directly solving fundamental matrix from dense
 optical flow correspondence and makes use of a two-view triangulation module to
 recover an up-to-scale 3D structure. Then, we align the scale of the depth pred
iction with the triangulated point cloud and use the transformed depth map for d
epth error computation and dense reprojection check. Our whole system can be joi
ntly trained end-to-end. Extensive experiments show that our system not only rea
ches state-of-the-art performance on KITTI depth and flow estimation, but also s
ignificantly improves the generalization ability of existing self-supervised dep
th-pose learning methods under a variety of challenging scenarios, and achieves
state-of-the-art results among self-supervised learning-based methods on KITTI O
dometry and NYUv2 dataset. Furthermore, we present some interesting findings on
the limitation of PoseNet-based relative pose estimation methods in terms of gen
eralization ability. Code is available at https://github.com/B1ueber2y/TrianFlow
.
****
Quasi-Newton Solver for Robust Non-Rigid Registration
Yuxin Yao, Bailin Deng, Weiwei Xu, Juyong Zhang; Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7600-760
9
Imperfect data (noise, outliers and partial overlap) and high degrees of freedom
 make non-rigid registration a classical challenging problem in computer vision.
 Existing methods typically adopt the l_p type robust estimator to regularize th
e fitting and smoothness, and the proximal operator is used to solve the resulti
ng non-smooth problem. However, the slow convergence of these algorithms limits
its wide applications. In this paper, we propose a formulation for robust non-ri
gid registration based on a globally smooth robust estimator for data fitting an
d regularization, which can handle outliers and partial overlaps. We apply the m
ajorization-minimization algorithm to the problem, which reduces each iteration
to solving a simple least-squares problem with L-BFGS. Extensive experiments dem
onstrate the effectiveness of our method for non-rigid alignment between two sha
pes with outliers and partial overlap. with quantitative evaluation showing that
 it outperforms state-of-the-art methods in terms of registration accuracy and c
omputational speed. The source code is available at https://github.com/Juyong/Fa
st_RNRR.
****
Multi-Scale Progressive Fusion Network for Single Image Deraining
Kui Jiang, Zhongyuan Wang, Peng Yi, Chen Chen, Baojin Huang, Yimin Luo, Ji
ayi Ma, Junjun Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision
 and Pattern Recognition (CVPR), 2020, pp. 8346-8355
Rain streaks in the air appear in various blurring degrees and resolutions due t
o different distances from their positions to the camera. Similar rain patterns

are visible in a rain image as well as its multi-scale (or multi-resolution) ver sions, which makes it possible to exploit such complementary information for rai n streak representation. In this work, we explore the multi-scale collaborative representation for rain streaks from the perspective of input image scales and h ierarchical deep features in a unified framework, termed multi-scale progressive fusion network (MSPFN) for single image rain streak removal. For the similar ra in streaks at different positions, we employ recurrent calculation to capture th e global texture, thus allowing to explore the complementary and redundant infor mation at the spatial dimension to characterize target rain streaks. Besides, we construct multi-scale pyramid structure, and further introduce the attention me chanism to guide the fine fusion of these correlated information from different scales. This multi-scale progressive fusion strategy not only promotes the coope rative representation, but also boosts the end-to-end training. Our proposed met hod is extensively evaluated on several benchmark datasets and achieves the stat e-of-the-art results. Moreover, we conduct experiments on joint deraining, detec tion, and segmentation tasks, and inspire a new research direction of vision tas k driven image deraining. The source code is available at https://github.com/kui hua/MSPFN.

****

Three-Dimensional Reconstruction of Human Interactions

Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, Cristian Sminchisescu; Proceedings of the IEEE/CVF Conference on Computer Visio n and Pattern Recognition (CVPR), 2020, pp. 7214-7223

Understanding 3d human interactions is fundamental for fine grained scene analys is and behavioural modeling. However, most of the existing models focus on analy zing a single person in isolation, and those who process several people focus la rgely on resolving multi-person data association, rather than inferring interact ions. This may lead to incorrect, lifeless 3d estimates, that miss the subtle hu man contact aspects--the essence of the event--and are of little use for detaile d behavioral understanding. This paper addresses such issues and makes several c ontributions: (1) we introduce models for interaction signature estimation (ISP) encompassing contact detection, segmentation, and 3d contact signature predicti on; (2) we show how such components can be leveraged in order to produce augment ed losses that ensure contact consistency during 3d reconstruction; (3) we const ruct several large datasets for learning and evaluating 3d contact prediction an d reconstruction methods; specifically, we introduce CHI3D, a lab-based accurate 3d motion capture dataset with 631 sequences containing 2,525 contact events, 7 28,664 ground truth 3d poses, as well as FlickrCI3D, a dataset of 11,216 images, with 14,081 processed pairs of people, and 81,233 facet-level surface correspon dences within 138,213 selected contact regions. Finally, (4) we present models a nd baselines to illustrate how contact estimation supports meaningful 3d reconst ruction where essential interactions are captured. Models and data are made avai lable for research purposes at http://vision.imar.ro/ci3d.

****

Real-Time Panoptic Segmentation From Dense Detections

Rui Hou, Jie Li, Arjun Bhargava, Allan Raventos, Vitor Guizilini, Chao Fang , Jerome Lynch, Adrien Gaidon; Proceedings of the IEEE/CVF Conference on Compu ter Vision and Pattern Recognition (CVPR), 2020, pp. 8523-8532

Panoptic segmentation is a complex full scene parsing task requiring simultaneou s instance and semantic segmentation at high resolution. Current state-of-the-ar t approaches cannot run in real-time, and simplifying these architectures to imp rove efficiency severely degrades their accuracy. In this paper, we propose a ne w single-shot panoptic segmentation network that leverages dense detections and a global self-attention mechanism to operate in real-time with performance appro aching the state of the art. We introduce a novel parameter-free mask constructi on method that substantially reduces computational complexity by efficiently reu sing information from the object detection and semantic segmentation sub-tasks. The resulting network has a simple data flow that requires no feature map re-sam pling, enabling significant hardware acceleration. Our experiments on the Citysc apes and COCO benchmarks show that our network works at 30 FPS on 1024x2048 reso

lution, trading a 3% relative performance degradation from the current state of the art for up to 440% faster inference.
****

NestedVAE: Isolating Common Factors via Weak Supervision
Matthew J. Vowels, Necati Cihan Camgoz, Richard Bowden; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9202-9212
Fair and unbiased machine learning is an important and active field of research, as decision processes are increasingly driven by models that learn from data. Unfortunately, any biases present in the data may be learned by the model, thereby inappropriately transferring that bias into the decision making process. We identify the connection between the task of bias reduction and that of isolating factors common between domains whilst encouraging domain specific invariance. To isolate the common factors we combine the theory of deep latent variable models with information bottleneck theory for scenarios whereby data may be naturally paired across domains and no additional supervision is required. The result is the Nested Variational AutoEncoder (NestedVAE). Two outer VAEs with shared weights attempt to reconstruct the input and infer a latent space, whilst a nested VAE attempts to reconstruct the latent representation of one image, from the latent representation of its paired image. In so doing, the nested VAE isolates the common latent factors/causes and becomes invariant to unwanted factors that are not shared between paired images. We also propose a new metric to provide a balanced method of evaluating consistency and classifier performance across domains which we refer to as the Adjusted Parity metric. An evaluation of NestedVAE on both domain and attribute invariance, change detection, and learning common factors for the prediction of biological sex demonstrates that NestedVAE significantly outperforms alternative methods.
****

Recurrent Feature Reasoning for Image Inpainting
Jingyuan Li, Ning Wang, Lefei Zhang, Bo Du, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7760-7768
Existing inpainting methods have achieved promising performance for recovering regular or small image defects. However, filling in large continuous holes remains difficult due to the lack of constraints for the hole center. In this paper, we devise a Recurrent Feature Reasoning (RFR) network which is mainly constructed by a plug-and-play Recurrent Feature Reasoning module and a Knowledge Consistent Attention (KCA) module. Analogous to how humans solve puzzles (i.e., first solve the easier parts and then use the results as additional information to solve difficult parts), the RFR module recurrently infers the hole boundaries of the convolutional feature maps and then uses them as clues for further inference. The module progressively strengthens the constraints for the hole center and the results become explicit. To capture information from distant places in the feature map for RFR, we further develop KCA and incorporate it in RFR. Empirically, we first compare the proposed RFR-Net with existing backbones, demonstrating that RFR-Net is more efficient (e.g., a 4% SSIM improvement for the same model size). We then place the network in the context of the current state-of-the-art, where it exhibits improved performance. The corresponding source code is available at: https://github.com/jingyuanli001/RFR-Inpainting
****

Harmonizing Transferability and Discriminability for Adapting Object Detectors
Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, Qi Dou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8869-8878
Recent advances in adaptive object detection have achieved compelling results in virtue of adversarial feature adaptation to mitigate the distributional shifts along the detection pipeline. Whilst adversarial adaptation significantly enhances the transferability of feature representations, the feature discriminability of object detectors remains less investigated. Moreover, transferability and discriminability may come at a contradiction in adversarial adaptation given the co

mplex combinations of objects and the differentiated scene layouts between domai
ns. In this paper, we propose a Hierarchical Transferability Calibration Network
 (HTCN) that hierarchically (local-region/image/instance) calibrates the transfe
rability of feature representations for harmonizing transferability and discrimi
nability. The proposed model consists of three components: (1) Importance Weight
ed Adversarial Training with input Interpolation (IWAT-I), which strengthens the
 global discriminability by re-weighting the interpolated image-level features;
(2) Context-aware Instance-Level Alignment (CILA) module, which enhances the loc
al discriminability by capturing the underlying complementary effect between the
 instance-level feature and the global context information for the instance-leve
l feature alignment; (3) local feature masks that calibrate the local transferab
ility to provide semantic guidance for the following discriminative pattern alig
nment. Experimental results show that HTCN significantly outperforms the state-o
f-the-art methods on benchmark datasets.
****

Unsupervised Magnification of Posture Deviations Across Subjects
Michael Dorkenwald, Uta Buchler, Bjorn Ommer; Proceedings of the IEEE/CVF Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8256-8266
Analyzing human posture and precisely comparing it across different subjects is
essential for accurate understanding of behavior and numerous vision application
s such as medical diagnostics, sports, or surveillance. Motion magnification tec
hniques help to see even small deviations in posture that are invisible to the n
aked eye. However, they fail when comparing subtle posture differences across in
dividuals with diverse appearance. Keypoint-based posture estimation and classif
ication techniques can handle large variations in appearance, but are invariant
to subtle deviations in posture. We present an approach to unsupervised magnific
ation of posture differences across individuals despite large deviations in appe
arance. We do not require keypoint annotation and visualize deviations on a sub-
bodypart level. To transfer appearance across subjects onto a magnified posture,
 we propose a novel loss for disentangling appearance and posture in an autoenco
der. Posture magnification yields exaggerated images that are different from the
 training set. Therefore, we incorporate magnification already into the training
 of the disentangled autoencoder and learn on real data and synthesized magnific
ations without supervision. Experiments confirm that our approach improves upon
the state-of-the-art in magnification and on the application of discovering post
ure deviations due to impairment.
****

PADS: Policy-Adapted Sampling for Visual Similarity Learning
Karsten Roth, Timo Milbich, Bjorn Ommer; Proceedings of the IEEE/CVF Conferenc
e on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6568-6577
Learning visual similarity requires to learn relations, typically between triple
ts of images. Albeit triplet approaches being powerful, their computational comp
lexity mostly limits training to only a subset of all possible training triplets
. Thus, sampling strategies that decide when to use which training sample during
 learning are crucial. Currently, the prominent paradigm are fixed or curriculum
 sampling strategies that are predefined before training starts. However, the pr
oblem truly calls for a sampling process that adjusts based on the actual state
of the similarity representation during training. We, therefore, employ reinforc
ement learning and have a teacher network adjust the sampling distribution based
 on the current state of the learner network, which represents visual similarity
. Experiments on benchmark datasets using standard triplet-based losses show tha
t our adaptive sampling strategy significantly outperforms fixed sampling strate
gies. Moreover, although our adaptive sampling is only applied on top of basic t
riplet-learning frameworks, we reach competitive results to state-of-the-art app
roaches that employ diverse additional learning signals or strong ensemble archi
tectures. Code can be found under https://github.com/Confusezius/CVPR2020_PADS.
****

Interactive Multi-Label CNN Learning With Partial Labels
Dat Huynh, Ehsan Elhamifar; Proceedings of the IEEE/CVF Conference on Computer
Vision and Pattern Recognition (CVPR), 2020, pp. 9423-9432

We address the problem of efficient end-to-end learning a multi-label Convolutional Neural Network (CNN) on training images with partial labels. Training a CNN with partial labels, hence a small number of images for every label, using the standard cross-entropy loss is prone to overfitting and performance drop. We introduce a new loss function that regularizes the cross-entropy loss with a cost function that measures the smoothness of labels and features of images on the data manifold. Given that optimizing the new loss function over the CNN parameters requires learning similarities among labels and images, which itself depends on knowing the parameters of the CNN, we develop an efficient interactive learning framework in which the two steps of similarity learning and CNN training interact and improve the performance of each another. Our method learns the CNN parameters without requiring keeping all training data in the memory, allows to learn few informative similarities only for images in each mini-batch and handles changing feature representations. By extensive experiments on Open Images, CUB and MS-COCO datasets,we demonstrate the effectiveness of our method. In particular, on the large-scale Open Images dataset, we improve the state of the art by 1.02% in mAP score over 5,000 classes.

****

SketchyCOCO: Image Generation From Freehand Scene Sketches

Chengying Gao, Qi Liu, Qi Xu, Limin Wang, Jianzhuang Liu, Changqing Zou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5174-5183

We introduce the first method for automatic image generation from scene-level freehand sketches. Our model allows for controllable image generation by specifying the synthesis goal via freehand sketches. The key contribution is an attribute vector bridged Generative Adversarial Network called EdgeGAN, which supports high visual-quality object-level image content generation without using freehand sketches as training data. We have built a large-scale composite dataset called SketchyCOCO to support and evaluate the solution. We validate our approach on the tasks of both object-level and scene-level image generation on SketchyCOCO. Through quantitative, qualitative results, human evaluation and ablation studies, we demonstrate the method's capacity to generate realistic complex scene-level images from various freehand sketches.

****

Effectively Unbiased FID and Inception Score and Where to Find Them

Min Jin Chong, David Forsyth; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6070-6079

This paper shows that two commonly used evaluation metrics for generative models, the Frechet Inception Distance (FID) and the Inception Score (IS), are biased -- the expected value of the score computed for a finite sample set is not the true value of the score. Worse, the paper shows that the bias term depends on the particular model being evaluated, so model A may get a better score than model B simply because model A's bias term is smaller. This effect cannot be fixed by evaluating at a fixed number of samples. This means all comparisons using FID or IS as currently computed are unreliable. We then show how to extrapolate the score to obtain an effectively bias-free estimate of scores computed with an infinite number of samples, which we term FID Infinity and IS Infinity. In turn, this effectively bias-free estimate requires good estimates of scores with a finite number of samples. We show that using Quasi-Monte Carlo integration notably improves estimates of FID and IS for finite sample sets. Our extrapolated scores are simple, drop-in replacements for the finite sample scores. Additionally, we show that using low discrepancy sequence in GAN training offers small improvements in the resulting generator.

****

Controllable Orthogonalization in Training DNNs

Lei Huang, Li Liu, Fan Zhu, Diwen Wan, Zehuan Yuan, Bo Li, Ling Shao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6429-6438

Orthogonality is widely used for training deep neural networks (DNNs) due to its ability to maintain all singular values of the Jacobian close to 1 and reduce r

edundancy in representation. This paper proposes a computationally efficient and numerically stable orthogonalization method using Newton's iteration (ONI), to learn a layer-wise orthogonal weight matrix in DNNs. ONI works by iteratively stretching the singular values of a weight matrix towards 1. This property enables it to control the orthogonality of a weight matrix by its number of iterations. We show that our method improves the performance of image classification networks by effectively controlling the orthogonality to provide an optimal tradeoff between optimization benefits and representational capacity reduction. We also show that ONI stabilizes the training of generative adversarial networks (GANs) by maintaining the Lipschitz continuity of a network, similar to spectral normalization (SN), and further outperforms SN by providing controllable orthogonality.
****
Interpreting the Latent Space of GANs for Semantic Face Editing
Yujun Shen, Jinjin Gu, Xiaoou Tang, Bolei Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9243-9252
Despite the recent advance of Generative Adversarial Networks (GANs) in high-fidelity image synthesis, there lacks enough understanding of how GANs are able to map a latent code sampled from a random distribution to a photo-realistic image. Previous work assumes the latent space learned by GANs follows a distributed representation but observes the vector arithmetic phenomenon. In this work, we propose a novel framework, called InterFaceGAN, for semantic face editing by interpreting the latent semantics learned by GANs. In this framework, we conduct a detailed study on how different semantics are encoded in the latent space of GANs for face synthesis. We find that the latent code of well-trained generative models actually learns a disentangled representation after linear transformations. We explore the disentanglement between various semantics and manage to decouple some entangled semantics with subspace projection, leading to more precise control of facial attributes. Besides manipulating gender, age, expression, and the presence of eyeglasses, we can even vary the face pose as well as fix the artifacts accidentally generated by GAN models. The proposed method is further applied to achieve real image manipulation when combined with GAN inversion methods or some encoder-involved models. Extensive results suggest that learning to synthesize faces spontaneously brings a disentangled and controllable facial attribute representation.
****
Tracking by Instance Detection: A Meta-Learning Approach
Guangting Wang, Chong Luo, Xiaoyan Sun, Zhiwei Xiong, Wenjun Zeng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6288-6297
We consider the tracking problem as a special type of object detection problem, which we call instance detection. With proper initialization, a detector can be quickly converted into a tracker by learning the new instance from a single image. We find that model-agnostic meta-learning (MAML) offers a strategy to initialize the detector that satisfies our needs. We propose a principled three-step approach to build a high-performance tracker. First, pick any modern object detector trained with gradient descent. Second, conduct offline training (or initialization) with MAML. Third, perform domain adaptation using the initial frame. We follow this procedure to build two trackers, named Retina-MAML and FCOS-MAML, based on two modern detectors RetinaNet and FCOS. Evaluations on four benchmarks show that both trackers are competitive against state-of-the-art trackers. On OTB-100, Retina-MAML achieves the highest ever AUC of 0.712. On TrackingNet, FCOS-MAML ranks the first on the leader board with an AUC of 0.757 and the normalized precision of 0.822. Both trackers run in real-time at 40 FPS.
****
Learned Image Compression With Discretized Gaussian Mixture Likelihoods and Attention Modules
Zhengxue Cheng, Heming Sun, Masaru Takeuchi, Jiro Katto; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7939-7948
Image compression is a fundamental research field and many well-known compressio

n standards have been developed for many decades. Recently, learned compression methods exhibit a fast development trend with promising results. However, there is still a performance gap between learned compression algorithms and reigning compression standards, especially in terms of widely used PSNR metric. In this paper, we explore the remaining redundancy of recent learned compression algorithms. We have found accurate entropy models for rate estimation largely affect the optimization of network parameters and thus affect the rate-distortion performance. Therefore, in this paper, we propose to use discretized Gaussian Mixture Likelihoods to parameterize the distributions of latent codes, which can achieve a more accurate and flexible entropy model. Besides, we take advantage of recent attention modules and incorporate them into network architecture to enhance the performance. Experimental results demonstrate our proposed method achieves a state-of-the-art performance compared to existing learned compression methods on both Kodak and high-resolution datasets. To our knowledge our approach is the first work to achieve comparable performance with latest compression standard Versatile Video Coding (VVC) regarding PSNR. More importantly, our approach generates more visually pleasant results when optimized by MS-SSIM.

****

PointGroup: Dual-Set Point Grouping for 3D Instance Segmentation

Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, Jiaya Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4867-4876

Instance segmentation is an important task for scene understanding. Compared to the fully-developed 2D, 3D instance segmentation for point clouds have much room to improve. In this paper, we present PointGroup, a new end-to-end bottom-up architecture, specifically focused on better grouping the points by exploring the void space between objects. We design a two-branch network to extract point features and predict semantic labels and offsets, for shifting each point towards its respective instance centroid. A clustering component is followed to utilize both the original and offset-shifted point coordinate sets, taking advantage of their complementary strength. Further, we formulate the ScoreNet to evaluate the candidate instances, followed by the Non-Maximum Suppression (NMS) to remove duplicates. We conduct extensive experiments on two challenging datasets, ScanNet v2 and S3DIS, on which our method achieves the highest performance, 63.6% and 64.0%, compared to 54.9% and 54.4% achieved by former best solutions in terms of mAP with IoU threshold 0.5.

****

Semantic Drift Compensation for Class-Incremental Learning

Lu Yu, Bartlomiej Twardowski, Xialei Liu, Luis Herranz, Kai Wang, Yongmei Cheng, Shangling Jui, Joost van de Weijer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6982-6991

Class-incremental learning of deep networks sequentially increases the number of classes to be classified. During training, the network has only access to data of one task at a time, where each task contains several classes. In this setting, networks suffer from catastrophic forgetting which refers to the drastic drop in performance on previous tasks. The vast majority of methods have studied this scenario for classification networks, where for each new task the classification layer of the network must be augmented with additional weights to make room for the newly added classes. Embedding networks have the advantage that new classes can be naturally included into the network without adding new weights. Therefore, we study incremental learning for embedding networks. In addition, we propose a new method to estimate the drift, called semantic drift, of features and compensate for it without the need of any exemplars. We approximate the drift of previous tasks based on the drift that is experienced by current task data. We perform experiments on fine-grained datasets, CIFAR100 and ImageNet-Subset. We demonstrate that embedding networks suffer significantly less from catastrophic forgetting. We outperform existing methods which do not require exemplars and obtain competitive results compared to methods which store exemplars. Furthermore, we show that our proposed SDC when combined with existing methods to prevent forgetting consistently improves results.

****
Generating 3D People in Scenes Without People
Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J. Black, Siyu Tang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6194-6204
We present a fully automatic system that takes a 3D scene and generates plausible 3D human bodies that are posed naturally in that 3D scene. Given a 3D scene without people, humans can easily imagine how people could interact with the scene and the objects in it. However, this is a challenging task for a computer as solving it requires that (1) the generated human bodies to be semantically plausible within the 3D environment (e.g. people sitting on the sofa or cooking near the stove), and (2) the generated human-scene interaction to be physically feasible such that the human body and scene do not interpenetrate while, at the same time, body-scene contact supports physical interactions. To that end, we make use of the surface-based 3D human model SMPL-X. We first train a conditional variational autoencoder to predict semantically plausible 3D human poses conditioned on latent scene representations, then we further refine the generated 3D bodies using scene constraints to enforce feasible physical interaction. We show that our approach is able to synthesize realistic and expressive 3D human bodies that naturally interact with 3D environment. We perform extensive experiments demonstrating that our generative framework compares favorably with existing methods, both qualitatively and quantitatively. We believe that our scene-conditioned 3D human generation pipeline will be useful for numerous applications; e.g. to generate training data for human pose estimation, in video games and in VR/AR. Our project page for data and code can be seen at: https://vlg.inf.ethz.ch/projects/PSI/ .
****
Computing Valid P-Values for Image Segmentation by Selective Inference
Kosuke Tanizaki, Noriaki Hashimoto, Yu Inatsu, Hidekata Hontani, Ichiro Takeuchi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9553-9562
Image segmentation is one of the most fundamental tasks in computer vision. In many practical applications, it is essential to properly evaluate the reliability of individual segmentation results. In this study, we propose a novel framework for quantifying the statistical significance of individual segmentation results in the form of p-values by statistically testing the difference between the object region and the background region. This seemingly simple problem is actually quite challenging because the difference --- called segmentation bias --- can be deceptively large due to the adaptation of the segmentation algorithm to the data. To overcome this difficulty, we introduce a statistical approach called selective inference, and develop a framework for computing valid p-values in which segmentation bias is properly accounted for. Although the proposed framework is potentially applicable to various segmentation algorithms, we focus in this paper on graph-cut- and threshold-based segmentation algorithms, and develop two specific methods for computing valid p-values for the segmentation results obtained by these algorithms. We prove the theoretical validity of these two methods and demonstrate their practicality by applying them to the segmentation of medical images.
****
Recursive Least-Squares Estimator-Aided Online Learning for Visual Tracking
Jin Gao, Weiming Hu, Yan Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7386-7395
Online learning is crucial to robust visual object tracking as it can provide high discrimination power in the presence of background distractors. However, there are two contradictory factors affecting its successful deployment on the real visual tracking platform: the discrimination issue due to the challenges in vanilla gradient descent, which does not guarantee good convergence; the robustness issue due to over-fitting resulting from excessive update with limited memory size (the oldest samples are discarded). Despite many dedicated techniques proposed to somehow treat those issues, in this paper we take a new way to strike a com

promise between them based on the recursive least-squares estimation (LSE) algorithm. After connecting each fully-connected layer with LSE separately via normal equations, we further propose an improved mini-batch stochastic gradient descent algorithm for fully-connected network learning with memory retention in a recursive fashion. This characteristic can spontaneously reduce the risk of over-fitting resulting from catastrophic forgetting in excessive online learning. Meanwhile, it can effectively improve convergence though the cost function is computed over all the training samples that the algorithm has ever seen. We realize this recursive LSE-aided online learning technique in the state-of-the-art RT-MDNet tracker, and the consistent improvements on four challenging benchmarks prove its efficiency without additional offline training and too much tedious work on parameter adjusting.

****

End-to-End Pseudo-LiDAR for Image-Based 3D Object Detection

Rui Qian, Divyansh Garg, Yan Wang, Yurong You, Serge Belongie, 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. 5881-5890

Reliable and accurate 3D object detection is a necessity for safe autonomous driving. Although LiDAR sensors can provide accurate 3D point cloud estimates of the environment, they are also prohibitively expensive for many settings. Recently, the introduction of pseudo-LiDAR (PL) has led to a drastic reduction in the accuracy gap between methods based on LiDAR sensors and those based on cheap stereo cameras. PL combines state-of-the-art deep neural networks for 3D depth estimation with those for 3D object detection by converting 2D depth map outputs to 3D point cloud inputs. However, so far these two networks have to be trained separately. In this paper, we introduce a new framework based on differentiable Change of Representation (CoR) modules that allow the entire PL pipeline to be trained end-to-end. The resulting framework is compatible with most state-of-the-art networks for both tasks and in combination with PointRCNN improves over PL consistently across all benchmarks --- yielding the highest entry on the KITTI image-based 3D object detection leaderboard at the time of submission. Our code will be made available at https://github.com/mileyan/pseudo-LiDAR_e2e.

****

A Quantum Computational Approach to Correspondence Problems on Point Sets

Vladislav Golyanik, Christian Theobalt; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9182-9191

Modern adiabatic quantum computers (AQC) are already used to solve difficult combinatorial optimisation problems in various domains of science. Currently, only a few applications of AQC in computer vision have been demonstrated. We review AQC and derive a new algorithm for correspondence problems on point sets suitable for execution on AQC. Our algorithm has a subquadratic computational complexity of the state preparation. Examples of successful transformation estimation and point set alignment by simulated sampling are shown in the systematic experimental evaluation. Finally, we analyse the differences in the solutions and the corresponding energy values.

****

High-Frequency Component Helps Explain the Generalization of Convolutional Neural Networks

Haohan Wang, Xindi Wu, Zeyi Huang, Eric P. Xing; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8684-8694

We investigate the relationship between the frequency spectrum of image data and the generalization behavior of convolutional neural networks (CNN). We first notice CNN's ability in capturing the high-frequency components of images. These high-frequency components are almost imperceptible to a human. Thus the observation leads to multiple hypotheses that are related to the generalization behaviors of CNN, including a potential explanation for adversarial examples, a discussion of CNN's trade-off between robustness and accuracy, and some evidence in understanding training heuristics.

****
Rotate-and-Render: Unsupervised Photorealistic Face Rotation From Single-View Images
Hang Zhou, Jihao Liu, Ziwei Liu, Yu Liu, Xiaogang Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5911-5920
Though face rotation has achieved rapid progress in recent years, the lack of high-quality paired training data remains a great hurdle for existing methods. The current generative models heavily rely on datasets with multi-view images of the same person. Thus, their generated results are restricted by the scale and domain of the data source. To overcome these challenges, we propose a novel unsupervised framework that can synthesize photo-realistic rotated faces using only single-view image collections in the wild. Our key insight is that rotating faces in the 3D space back and forth, and re-rendering them to the 2D plane can serve as a strong self-supervision. We leverage the recent advances in 3D face modeling and high-resolution GAN to constitute our building blocks. Since the 3D rotation-and-render on faces can be applied to arbitrary angles without losing details, our approach is extremely suitable for in-the-wild scenarios (i.e. no paired data are available), where existing methods fall short. Extensive experiments demonstrate that our approach has superior synthesis quality as well as identity preservation over the state-of-the-art methods, across a wide range of poses and domains. Furthermore, we validate that our rotate-and-render framework naturally can act as an effective data augmentation engine for boosting modern face recognition systems even on strong baseline models
****
Scene-Adaptive Video Frame Interpolation via Meta-Learning
Myungsub Choi, Janghoon Choi, Sungyong Baik, Tae Hyun Kim, Kyoung Mu Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9444-9453
Video frame interpolation is a challenging problem because there are different scenarios for each video depending on the variety of foreground and background motion, frame rate, and occlusion. It is therefore difficult for a single network with fixed parameters to generalize across different videos. Ideally, one could have a different network for each scenario, but this is computationally infeasible for practical applications. In this work, we propose to adapt the model to each video by making use of additional information that is readily available at test time and yet has not been exploited in previous works. We first show the benefits of 'test-time adaptation' through simple fine-tuning of a network, then we greatly improve its efficiency by incorporating meta-learning. We obtain significant performance gains with only a single gradient update without any additional parameters. Finally, we show that our meta-learning framework can be easily employed to any video frame interpolation network and can consistently improve its performance on multiple benchmark datasets.
****
On the Distribution of Minima in Intrinsic-Metric Rotation Averaging
Kyle Wilson, David Bindel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6031-6039
Rotation Averaging is a non-convex optimization problem that determines orientations of a collection of cameras from their images of a 3D scene. The problem has been studied using a variety of distances and robustifiers. The intrinsic (or geodesic) distance on SO(3) is geometrically meaningful; but while some extrinsic distance-based solvers admit (conditional) guarantees of correctness, no comparable results have been found under the intrinsic metric. In this paper, we study the spatial distribution of local minima. First, we do a novel empirical study to demonstrate sharp transitions in qualitative behavior: as problems become noisier, they transition from a single (easy-to-find) dominant minimum to a cost surface filled with minima. In the second part of this paper we derive a theoretical bound for when this transition occurs. This is an extension of the results of [24], which used local convexity as a proxy to study the difficulty of problem. By recognizing the underly- ing quotient manifold geometry of the problem we ac

hieve an n-fold improvement over prior work. Incidentally, our analysis also ext
ends the prior l2 work to general lp costs. Our results suggest using algebraic
connectivity as an indicator of problem difficulty.
****

Explainable Object-Induced Action Decision for Autonomous Vehicles
Yiran Xu, Xiaoyin Yang, Lihang Gong, Hsuan-Chu Lin, Tz-Ying Wu, Yunsheng Li
, Nuno Vasconcelos; Proceedings of the IEEE/CVF Conference on Computer Vision a
nd Pattern Recognition (CVPR), 2020, pp. 9523-9532
A new paradigm is proposed for autonomous driving. The new paradigm lies between
the end-to-end and pipelined approaches, and is inspired by how humans solve th
e problem. While it relies on scene understanding, the latter only considers obj
ects that could originate hazard. These are denoted as action inducing, since ch
anges in their state should trigger vehicle actions. They also define a set of e
xplanations for these actions, which should be produced jointly with the latter.
An extension of the BDD100K dataset, annotated for a set of 4 actions and 21 ex
planations, is proposed. A new multi-task formulation of the problem, which opti
mizes the accuracy of both action commands and explanations, is then introduced.
A CNN architecture is finally proposed to solve this problem, by combining reas
oning about action inducing objects and global scene context. Experimental resul
ts show that the requirement of explanations improves the recognition of action-
inducing objects, which in turn leads to better action predictions.
****

DLWL: Improving Detection for Lowshot Classes With Weakly Labelled Data
Vignesh Ramanathan, Rui Wang, Dhruv Mahajan; Proceedings of the IEEE/CVF Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9342-9352
Large detection datasets have a long tail of lowshot classes with very few bound
ing box annotations. We wish to improve detection for lowshot classes with weakl
y labelled web-scale datasets only having image-level labels. This requires a de
tection framework that can be jointly trained with limited number of bounding bo
x annotated images and large number of weakly labelled images. Towards this end,
we propose a modification to the FRCNN model to automatically infer label assig
nment for objects proposals from weakly labelled images during training. We pose
this label assignment as a Linear Program with constraints on the number and ov
erlap of object instances in an image. We show that this can be solved efficient
ly during training for weakly labelled images. Compared to just training with fe
w annotated examples, augmenting with weakly labelled examples in our framework
provides significant gains. We demonstrate this on the LVIS dataset 3.5 gain in
AP as well as different lowshot variants of the COCO dataset. We provide a thoro
ugh analysis of the effect of amount of weakly labelled and fully labelled data
required to train the detection model. Our DLWL framework can also outperform se
lf-supervised baselines like omni-supervision for lowshot classes.
****

Robust Partial Matching for Person Search in the Wild
Yingji Zhong, Xiaoyu Wang, Shiliang Zhang; Proceedings of the IEEE/CVF Confere
nce on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6827-6835
Various factors like occlusions, backgrounds, etc., would lead to misaligned det
ected bounding boxes , e.g., ones covering only portions of human body. This iss
ue is common but overlooked by previous person search works. To alleviate this i
ssue, this paper proposes an Align-to-Part Network (APNet) for person detection
and re-Identification (reID). APNet refines detected bounding boxes to cover the
estimated holistic body regions, from which discriminative part features can be
extracted and aligned. Aligned part features naturally formulate reID as a part
ial feature matching procedure, where valid part features are selected for simil
arity computation, while part features on occluded or noisy regions are discarde
d. This design enhances the robustness of person search to real-world challenges
with marginal computation overhead. This paper also contributes a Large-Scale d
ataset for Person Search in the wild (LSPS), which is by far the largest and the
most challenging dataset for person search. Experiments show that APNet brings
considerable performance improvement on LSPS. Meanwhile, it achieves competitive
performance on existing person search benchmarks like CUHK-SYSU and PRW.

****
Watch Your Up-Convolution: CNN Based Generative Deep Neural Networks Are Failing to Reproduce Spectral Distributions
Ricard Durall, Margret Keuper, Janis Keuper; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7890-7899
Generative convolutional deep neural networks, e.g. popular GAN architectures, are relying on convolution based up-sampling methods to produce non-scalar outputs like images or video sequences. In this paper, we show that common up-sampling methods, i.e. known as up-convolution or transposed convolution, are causing the inability of such models to reproduce spectral distributions of natural training data correctly. This effect is independent of the underlying architecture and we show that it can be used to easily detect generated data like deepfakes with up to 100% accuracy on public benchmarks. To overcome this drawback of current generative models, we propose to add a novel spectral regularization term to the training optimization objective. We show that this approach not only allows to train spectral consistent GANs that are avoiding high frequency errors. Also, we show that a correct approximation of the frequency spectrum has positive effects on the training stability and output quality of generative networks.
****
The Devil Is in the Details: Delving Into Unbiased Data Processing for Human Pose Estimation
Junjie Huang, Zheng Zhu, Feng Guo, Guan Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5700-5709
Recently, the leading performance of human pose estimation is dominated by top-down methods. Being a fundamental component in training and inference, data processing has not been systematically considered in pose estimation community, to the best of our knowledge. In this paper, we focus on this problem and find that the devil of top-down pose estimator is in the biased data processing. Specifically, by investigating the standard data processing in state-of-the-art approaches mainly including data transformation and encoding-decoding, we find that the results obtained by common flipping strategy are unaligned with the original ones in inference. Moreover, there is statistical error in standard encoding-decoding during both training and inference. Two problems couple together and significantly degrade the pose estimation performance. Based on quantitative analyses, we then formulate a principled way to tackle this dilemma. Data is processed in continuous space based on unit length (the intervals between pixels) instead of in discrete space with pixel, and a combined classification and regression approach is adopted to perform encoding-decoding. The Unbiased Data Processing (UDP) for human pose estimation can be achieved by combining the two together. UDP not only boosts the performance of existing methods by a large margin but also plays a important role in result reproducing and future exploration. As a model-agnostic approach, UDP promotes SimpleBaseline-ResNet50-256x192 by 1.5 AP (70.2 to 71.7) and HRNet-W32-256x192 by 1.7 AP (73.5 to 75.2) on COCO test-dev set. The HRNet-W48-384x288 equipped with UDP achieves 76.5 AP and sets a new state-of-the-art for human pose estimation. The source code is publicly available for further research.
****
GraphTER: Unsupervised Learning of Graph Transformation Equivariant Representations via Auto-Encoding Node-Wise Transformations
Xiang Gao, Wei Hu, Guo-Jun Qi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7163-7172
Recent advances in Graph Convolutional Neural Networks (GCNNs) have shown their efficiency for nonEuclidean data on graphs, which often require a large amount of labeled data with high cost. It it thus critical to learn graph feature representations in an unsupervised manner in practice. To this end, we propose a novel unsupervised learning of Graph Transformation Equivariant Representations (GraphTER), aiming to capture intrinsic patterns of graph structure under both global and local transformations. Specifically, we allow to sample different groups of nodes from a graph and then transform them node-wise isotropically or anisotropically. Then, we self-train a representation encoder to capture the graph struct

ures by reconstructing these node-wise transformations from the feature representations of the original and transformed graphs. In experiments, we apply the learned GraphTER to graphs of 3D point cloud data, and results on point cloud segmentation/classification show that GraphTER significantly outperforms state-of-the-art unsupervised approaches and pushes greatly closer towards the upper bound set by the fully supervised counterparts. The code is available at: https://github.com/gyshgx868/graph-ter.

****

UC-Net: Uncertainty Inspired RGB-D Saliency Detection via Conditional Variational Autoencoders

Jing Zhang, Deng-Ping Fan, Yuchao Dai, Saeed Anwar, Fatemeh Sadat Saleh, Tong Zhang, Nick Barnes; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8582-8591

In this paper, we propose the first framework (UCNet) to employ uncertainty for RGB-D saliency detection by learning from the data labeling process. Existing RGB-D saliency detection methods treat the saliency detection task as a point estimation problem, and produce a single saliency map following a deterministic learning pipeline. Inspired by the saliency data labeling process, we propose probabilistic RGB-D saliency detection network via conditional variational autoencoders to model human annotation uncertainty and generate multiple saliency maps for each input image by sampling in the latent space. With the proposed saliency consensus process, we are able to generate an accurate saliency map based on these multiple predictions. Quantitative and qualitative evaluations on six challenging benchmark datasets against 18 competing algorithms demonstrate the effectiveness of our approach in learning the distribution of saliency maps, leading to a new state-of-the-art in RGB-D saliency detection.

****

4D Visualization of Dynamic Events From Unconstrained Multi-View Videos

Aayush Bansal, Minh Vo, Yaser Sheikh, Deva Ramanan, Srinivasa Narasimhan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5366-5375

We present a data-driven approach for 4D space-time visualization of dynamic events from videos captured by hand-held multiple cameras. Key to our approach is the use of self-supervised neural networks specific to the scene to compose static and dynamic aspects of an event. Though captured from discrete viewpoints, this model enables us to move around the space-time of the event continuously. This model allows us to create virtual cameras that facilitate: (1) freezing the time and exploring views; (2) freezing a view and moving through time; and (3) simultaneously changing both time and view. We can also edit the videos and reveal occluded objects for a given view if it is visible in any of the other views. We validate our approach on challenging in-the-wild events captured using up to 15 mobile cameras.

****

Factorized Higher-Order CNNs With an Application to Spatio-Temporal Emotion Estimation

Jean Kossaifi, Antoine Toisoul, Adrian Bulat, Yannis Panagakis, Timothy M. Hospedales, Maja Pantic; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6060-6069

Training deep neural networks with spatio-temporal (i.e., 3D) or multidimensional convolutions of higher-order is computationally challenging due to millions of unknown parameters across dozens of layers. To alleviate this, one approach is to apply low-rank tensor decompositions to convolution kernels in order to compress the network and reduce its number of parameters. Alternatively, new convolutional blocks, such as MobileNet, can be directly designed for efficiency. In this paper, we unify these two approaches by proposing a tensor factorization framework for efficient multidimensional (separable) convolutions of higher-order. Interestingly, the proposed framework enables a novel higher-order transduction, allowing to train a network on a given domain (e.g., 2D images or N-dimensional data in general) and using transduction to generalize to higher-order data such as videos (or (N+K)--dimensional data in general), capturing for instance tempora

l dynamics while preserving the learnt spatial information. We apply the propose d methodology, coined CP-Higher-Order Convolution (HO-CPConv), to spatio-tempora l facial emotion analysis. Most existing facial affect models focus on static im agery and discard all temporal information. This is due to the above-mentioned b urden of training 3D convolutional nets and the lack of large bodies of video da ta annotated by experts. We address both issues with our proposed framework. Ini tial training is first done on static imagery before using transduction to gener alize to the temporal domain. We demonstrate superior performance on three chall enging large scale affect estimation datasets, AffectNet, SEWA, and AFEW-VA.
****

Hardware-in-the-Loop End-to-End Optimization of Camera Image Processing Pipeline s

Ali Mosleh, Avinash Sharma, Emmanuel Onzon, Fahim Mannan, Nicolas Robidoux, Felix Heide; Proceedings of the IEEE/CVF Conference on Computer Vision and Patt ern Recognition (CVPR), 2020, pp. 7529-7538

Commodity imaging systems rely on hardware image signal processing (ISP) pipelin es. These low-level pipelines consist of a sequence of processing blocks that, d epending on their hyperparameters, reconstruct a color image from RAW sensor mea surements. Hardware ISP hyperparameters have a complex interaction with the outp ut image, and therefore with the downstream application ingesting these images. Traditionally, ISPs are manually tuned in isolation by imaging experts without a n end-to-end objective. Very recently, ISPs have been optimized with 1st-order m ethods that require differentiable approximations of the hardware ISP. Departing from such approximations, we present a hardware-in-the-loop method that directl y optimizes hardware image processing pipelines for end-to-end domain-specific l osses by solving a nonlinear multi-objective optimization problem with a novel 0 th-order stochastic solver directly interfaced with the hardware ISP. We validat e the proposed method with recent hardware ISPs and 2D object detection, segment ation, and human viewing as end-to-end downstream tasks. For automotive 2D objec t detection, the proposed method outperforms manual expert tuning by 30% mean av erage precision (mAP) and recent methods using ISP approximations by 18% mAP.
****

VOLDOR: Visual Odometry From Log-Logistic Dense Optical Flow Residuals

Zhixiang Min, Yiding Yang, Enrique Dunn; Proceedings of the IEEE/CVF Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4898-4909

We propose a dense indirect visual odometry method taking as input externally es timated optical flow fields instead of hand-crafted feature correspondences. We define our problem as a probabilistic model and develop a generalized-EM formula tion for the joint inference of camera motion, pixel depth, and motion-track con fidence. Contrary to traditional methods assuming Gaussian-distributed observati on errors, we supervise our inference framework under an (empirically validated) adaptive log-logistic distribution model. Moreover, the log-logistic residual m odel generalizes well to different state-of-the-art optical flow methods, making our approach modular and agnostic to the choice of optical flow estimators. Our method achieved top-ranking results on both TUM RGB-D and KITTI odometry benchm arks. Our open-sourced implementation is inherently GPU-friendly with only linea r computational and storage growth.
****

P2B: Point-to-Box Network for 3D Object Tracking in Point Clouds

Haozhe Qi, Chen Feng, Zhiguo Cao, Feng Zhao, Yang Xiao; Proceedings of the I EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6329-6338

Towards 3D object tracking in point clouds, a novel point-to-box network termed P2B is proposed in an end-to-end learning manner. Our main idea is to first loca lize potential target centers in 3D search area embedded with target information . Then point-driven 3D target proposal and verification are executed jointly. In this way, the time-consuming 3D exhaustive search can be avoided. Specifically, we first sample seeds from the point clouds in template and search area respect ively. Then, we execute permutation-invariant feature augmentation to embed targ et clues from template into search area seeds and represent them with target-spe

cific features. Consequently, the augmented search area seeds regress the potential target centers via Hough voting. The centers are further strengthened with seed-wise targetness scores. Finally, each center clusters its neighbors to leverage the ensemble power for joint 3D target proposal and verification. We apply PointNet++ as our backbone and experiments on KITTI tracking dataset demonstrate P2B's superiority (10%'s improvement over state-of-the-art). Note that P2B can run with 40FPS on a single NVIDIA 1080Ti GPU. Our code and model are available at https://github.com/HaozheQi/P2B.
****

Unsupervised Deep Shape Descriptor With Point Distribution Learning
Yi Shi, Mengchen Xu, Shuaihang Yuan, Yi Fang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9353-9362
Deep learning models have achieved great success in supervised shape descriptor learning for 3D shape retrieval, classification, and correspondence. However, the unsupervised shape descriptor calculated via deep learning is less studied than that of supervised ones due to the design challenges of unsupervised neural network architecture. This paper proposes a novel probabilistic framework for the learning of unsupervised deep shape descriptors with point distribution learning. In our approach, we firstly associate each point with a Gaussian, and the point clouds are modeled as the distribution of the points. We then use deep neural networks (DNNs) to model a maximum likelihood estimation process that is traditionally solved with an iterative Expectation-Maximization (EM) process. Our key novelty is that "training" these DNNs with unsupervised self-correspondence L2 distance loss will elegantly reveal the statically significant deep shape descriptor representation for the distribution of the point clouds. We have conducted experiments over various 3D datasets. Qualitative and quantitative comparisons demonstrate that our proposed method achieves superior classification performance over existing unsupervised 3D shape descriptors. In addition, we verified the following attractive properties of our shape descriptor through experiments: multi-scale shape representation, robustness to shape rotation, and robustness to noise.
****

Learning to Transfer Texture From Clothing Images to 3D Humans
Aymen Mir, Thiemo Alldieck, Gerard Pons-Moll; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7023-7034
In this paper, we present a simple yet effective method to automatically transfer textures of clothing images (front and back) to 3D garments worn on top SMPL, in real time. We first automatically compute training pairs of images with aligned 3D garments using a custom non-rigid 3D to 2D registration method, which is accurate but slow. Using these pairs, we learn a mapping from pixels to the 3D garment surface. Our idea is to learn dense correspondences from garment image silhouettes to a 2D-UV map of a 3D garment surface using shape information alone, completely ignoring texture, which allows us to generalize to the wide range of web images. Several experiments demonstrate that our model is more accurate than widely used baselines such as thin-plate-spline warping and image-to-image translation networks while being orders of magnitude faster. Our model opens the door for applications such as virtual try-on, and allows for generation of 3D humans with varied textures which is necessary for learning. Code will be available at https://virtualhumans.mpi-inf.mpg.de/pix2surf/.
****

Disentangled and Controllable Face Image Generation via 3D Imitative-Contrastive Learning
Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, Xin Tong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5154-5163
We propose an approach for face image generation of virtual people with disentangled, precisely-controllable latent representations for identity of non-existing people, expression, pose, and illumination. We embed 3D priors into adversarial learning and train the network to imitate the image formation of an analytic 3D face deformation and rendering process. To deal with the generation freedom ind

uced by the domain gap between real and rendered faces, we further introduce con
trastive learning to promote disentanglement by comparing pairs of generated ima
ges. Experiments show that through our imitative-contrastive learning, the facto
r variations are very well disentangled and the properties of a generated face c
an be precisely controlled. We also analyze the learned latent space and present
 several meaningful properties supporting factor disentanglement. Our method can
 also be used to embed real images into the disentangled latent space. We hope o
ur method could provide new understandings of the relationship between physical
properties and deep image synthesis.
****

Multi-Scale Interactive Network for Salient Object Detection
Youwei Pang, Xiaoqi Zhao, Lihe Zhang, Huchuan Lu; Proceedings of the IEEE/CVF
 Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9413-94
22
Deep-learning based salient object detection methods achieve great progress. How
ever, the variable scale and unknown category of salient objects are great chall
enges all the time. These are closely related to the utilization of multi-level
and multi-scale features. In this paper, we propose the aggregate interaction mo
dules to integrate the features from adjacent levels, in which less noise is int
roduced because of only using small up-/down-sampling rates. To obtain more effi
cient multi-scale features from the integrated features, the self-interaction mo
dules are embedded in each decoder unit. Besides, the class imbalance issue caus
ed by the scale variation weakens the effect of the binary cross entropy loss an
d results in the spatial inconsistency of the predictions. Therefore, we exploit
 the consistency-enhanced loss to highlight the fore-/back-ground difference and
 preserve the intra-class consistency. Experimental results on five benchmark da
tasets demonstrate that the proposed method without any post-processing performs
 favorably against 23 state-of-the-art approaches. The source code will be publi
cly available at https://github.com/lartpang/MINet.
****

Correlation-Guided Attention for Corner Detection Based Visual Tracking
Fei Du, Peng Liu, Wei Zhao, Xianglong Tang; Proceedings of the IEEE/CVF Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6836-6845
Accurate bounding box estimation has recently attracted much attention in the tr
acking community because traditional multi-scale search strategies cannot estima
te tight bounding boxes in many challenging scenarios involving changes to the t
arget. A tracker capable of detecting target corners can flexibly adapt to such
changes, but existing corner detection based tracking methods have not achieved
adequate success. We analyze the reasons for their failure and propose a state-o
f-the-art tracker that performs correlation-guided attentional corner detection
in two stages. First, a region of interest (RoI) is obtained by employing an eff
icient Siamese network to distinguish the target from the background. Second, a
pixel-wise correlation-guided spatial attention module and a channel-wise correl
ation-guided channel attention module exploit the relationship between the targe
t template and the RoI to highlight corner regions and enhance features of the R
oI for corner detection. The correlation-guided attention modules improve the ac
curacy of corner detection, thus enabling accurate bounding box estimation. When
 trained on large-scale datasets using a novel RoI augmentation strategy, the pe
rformance of the proposed tracker, running at a high speed of 70 FPS, is compara
ble with that of state-of-the-art trackers in meeting five challenging performan
ce benchmarks.
****

Accurate Estimation of Body Height From a Single Depth Image via a Four-Stage De
veloping Network
Fukun Yin, Shizhe Zhou; Proceedings of the IEEE/CVF Conference on Computer Visi
on and Pattern Recognition (CVPR), 2020, pp. 8267-8276
Non-contact measurement of human body height can be very difficult under some ci
rcumstances.In this paper we address the problem of accurately estimating the he
ight of a person with arbitrary postures from a single depth image. By introduci
ng a novel part-based intermediate representation plus a four-stage increasingly

complex deep neural network, we manage to achieve significantly higher accuracy than previous methods. We first describe the human body in the form of a segmentation of human torso as four nearly rigid parts and then predict their lengths respectively by 3 CNNs. Instead of directly adding the lengths of these parts together, we further construct another independent developing CNN that combines the intermediate representation, part lengths and depth information together to finally predict the body height results.Here we develop an increasingly complex network architecture and adopt a hybrid pooling to optimize training process. To the best of our knowledge, this is the first method that estimates height only from a single depth image. In experiments our average accuracy reaches at 99.1% for people in various positions and postures.

****

AD-Cluster: Augmented Discriminative Clustering for Domain Adaptive Person Re-Identification

Yunpeng Zhai, Shijian Lu, Qixiang Ye, Xuebo Shan, Jie Chen, Rongrong Ji, Yonghong Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9021-9030

Domain adaptive person re-identification (re-ID) is a challenging task, especially when person identities in target domains are unknown. Existing methods attempt to address this challenge by transferring image styles or aligning feature distributions across domains, whereas the rich unlabeled samples in target domains are not sufficiently exploited. This paper presents a novel augmented discriminative clustering (AD-Cluster) technique that estimates and augments person clusters in target domains and enforces the discrimination ability of re-ID models with the augmented clusters. AD-Cluster is trained by iterative density-based clustering, adaptive sample augmentation, and discriminative feature learning. It learns an image generator and a feature encoder which aim to maximize the intra-cluster diversity in the sample space and minimize the intra-cluster distance in the feature space in an adversarial min-max manner. Finally, AD-Cluster increases the diversity of sample clusters and improves the discrimination capability of re-ID models greatly. Extensive experiments over Market-1501 and DukeMTMC-reID show that AD-Cluster outperforms the state-of-the-art with large margins.

****

Regularizing Neural Networks via Minimizing Hyperspherical Energy

Rongmei Lin, Weiyang Liu, Zhen Liu, Chen Feng, Zhiding Yu, James M. Rehg, Li Xiong, Le Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6917-6927

Inspired by the Thomson problem in physics where the distribution of multiple propelling electrons on a unit sphere can be modeled via minimizing some potential energy, hyperspherical energy minimization has demonstrated its potential in regularizing neural networks and improving their generalization power. In this paper, we first study the important role that hyperspherical energy plays in neural network training by analyzing its training dynamics. Then we show that naively minimizing hyperspherical energy suffers from some difficulties due to highly non-linear and non-convex optimization as the space dimensionality becomes higher, therefore limiting the potential to further improve the generalization. To address these problems, we propose the compressive minimum hyperspherical energy (CoMHE) as a more effective regularization for neural networks. Specifically, CoMHE utilizes projection mappings to reduce the dimensionality of neurons and minimizes their hyperspherical energy. According to different designs for the projection mapping, we propose several distinct yet well-performing variants and provide some theoretical guarantees to justify their effectiveness. Our experiments show that CoMHE consistently outperforms existing regularization methods, and can be easily applied to different neural networks.

****

Density-Aware Feature Embedding for Face Clustering

Senhui Guo, Jing Xu, Dapeng Chen, Chao Zhang, Xiaogang Wang, Rui Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6698-6706

Clustering has many applications in research and industry. However, traditional

clustering methods, such as K-means, DBSCAN and HAC, impose oversimplifying assu mptions and thus are not well-suited to face clustering. To adapt to the distrib ution of realistic problems, a natural approach is to use Graph Convolutional Ne tworks (GCNs) to enhance features for clustering. However, GCNs can only utilize local information, which ignores the overall characterisitcs of the clusters. I n this paper, we propose a Density-Aware Feature Embedding Network (DA-Net) for the task of face clustering, which utilizes both local and non-local information , to learn a robust feature embedding. Specifically, DA-Net uses GCNs to aggrega te features locally, and then incorporates non-local information using a density chain, which is a chain of faces from low density to high density. This density chain exploits the non-uniform distribution of face images in the dataset. Then , an LSTM takes the density chain as input to generate the final feature embeddi ng. Once this embedding is generated, traditional clustering methods, such as de nsity-based clustering, can be used to obtain the final clustering results. Exte nsive experiments verify the effectiveness of the proposed feature embedding met hod, which can achieve state-of-the-art performance on public benchmarks.
****

Learning to Manipulate Individual Objects in an Image
Yanchao Yang, Yutong Chen, Stefano Soatto; Proceedings of the IEEE/CVF Confere nce on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6558-6567
We describe a method to train a generative model with latent factors that are (a pproximately) independent and localized. This means that perturbing the latent v ariables affects only local regions of the synthesized image, corresponding to o bjects. Unlike other unsupervised generative models, ours enables object-centric manipulation, without requiring object-level annotations, or any form of annota tion for that matter. The key to our method is the combination of spatial disent anglement, enforced by a Contextual Information Separation loss, and perceptual cycle-consistency, enforced by a loss that penalizes changes in the image partit ion in response to perturbations of the latent factors. We test our method's abi lity to allow independent control of spatial and semantic factors of variability on existing datasets and also introduce two new ones that highlight the limitat ions of current methods.
****

3D Photography Using Context-Aware Layered Depth Inpainting
Meng-Li Shih, Shih-Yang Su, Johannes Kopf, Jia-Bin Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8028-8038
We propose a method for converting a single RGB-D input image into a 3D photo, i .e., a multi-layer representation for novel view synthesis that contains halluci nated color and depth structures in regions occluded in the original view. We us e a Layered Depth Image with explicit pixel connectivity as underlying represent ation, and present a learning-based inpainting model that iteratively synthesize s new local color-and-depth content into the occluded region in a spatial contex t-aware manner. The resulting 3D photos can be efficiently rendered with motion parallax using standard graphics engines. We validate the effectiveness of our m ethod on a wide range of challenging everyday scenes and show less artifacts whe n compared with the state-of-the-arts.
****

Grid-GCN for Fast and Scalable Point Cloud Learning
Qiangeng Xu, Xudong Sun, Cho-Ying Wu, Panqu Wang, Ulrich Neumann; Proceeding s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5661-5670
Due to the sparsity and irregularity of the point cloud data, methods that direc tly consume points have become popular. Among all point-based models, graph conv olutional networks (GCN) lead to notable performance by fully preserving the dat a granularity and exploiting point interrelation. However, point-based networks spend a significant amount of time on data structuring (e.g., Farthest Point Sam pling (FPS) and neighbor points querying), which limit the speed and scalability . In this paper, we present a method, named Grid-GCN, for fast and scalable poin t cloud learning. Grid-GCN uses a novel data structuring strategy, Coverage-Awar

e Grid Query (CAGQ). By leveraging the efficiency of grid space, CAGQ improves s patial coverage while reducing the theoretical time complexity. Compared with po pular sampling methods such as Farthest Point Sampling (FPS) and Ball Query, CAG Q achieves up to 50 times speed-up. With a Grid Context Aggregation (GCA) module , Grid-GCN achieves state-of-the-art performance on major point cloud classifica tion and segmentation benchmarks with significantly faster runtime than previous studies. Remarkably, Grid-GCN achieves the inference speed of 50FPS on ScanNet using 81920 points as input. The supplementary xharlie.github.io/papers/GGCN_sup CamReady.pdf and the code github.com/xharlie/Grid-GCN are released.
****

KFNet: Learning Temporal Camera Relocalization Using Kalman Filtering
Lei Zhou, Zixin Luo, Tianwei Shen, Jiahui Zhang, Mingmin Zhen, Yao Yao, Ti an Fang, Long Quan; Proceedings of the IEEE/CVF Conference on Computer Vision a nd Pattern Recognition (CVPR), 2020, pp. 4919-4928
Temporal camera relocalization estimates the pose with respect to each video fra me in sequence, as opposed to one-shot relocalization which focuses on a still i mage. Even though the time dependency has been taken into account, current tempo ral relocalization methods still generally underperform the state-of-the-art one -shot approaches in terms of accuracy. In this work, we improve the temporal rel ocalization method by using a network architecture that incorporates Kalman filt ering (KFNet) for online camera relocalization. In particular, KFNet extends the scene coordinate regression problem to the time domain in order to recursively establish 2D and 3D correspondences for the pose determination. The network arch itecture design and the loss formulation are based on Kalman filtering in the co ntext of Bayesian learning. Extensive experiments on multiple relocalization ben chmarks demonstrate the high accuracy of KFNet at the top of both one-shot and t emporal relocalization approaches.
****

SuperGlue: Learning Feature Matching With Graph Neural Networks
Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, Andrew Rabinovich; Pr oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4938-4947
This paper introduces SuperGlue, a neural network that matches two sets of local features by jointly finding correspondences and rejecting non-matchable points. Assignments are estimated by solving a differentiable optimal transport problem , whose costs are predicted by a graph neural network. We introduce a flexible c ontext aggregation mechanism based on attention, enabling SuperGlue to reason ab out the underlying 3D scene and feature assignments jointly. Compared to traditi onal, hand-designed heuristics, our technique learns priors over geometric trans formations and regularities of the 3D world through end-to-end training from ima ge pairs. SuperGlue outperforms other learned approaches and achieves state-of-t he-art results on the task of pose estimation in challenging real-world indoor a nd outdoor environments. The proposed method performs matching in real-time on a modern GPU and can be readily integrated into modern SfM or SLAM systems. The c ode and trained weights are publicly available at github.com/magicleap/SuperGlue PretrainedNetwork.
****

Probabilistic Structural Latent Representation for Unsupervised Embedding
Mang Ye, Jianbing Shen; Proceedings of the IEEE/CVF Conference on Computer Visi on and Pattern Recognition (CVPR), 2020, pp. 5457-5466
Unsupervised embedding learning aims at extracting low-dimensional visually mean ingful representations from large-scale unlabeled images, which can then be dire ctly used for similarity-based search. This task faces two major challenges: 1) mining positive supervision from highly similar fine-grained classes and 2) gene rating to unseen testing categories. To tackle these issues, this paper proposes a probabilistic structural latent representation (PSLR), which incorporates an adaptable softmax embedding to approximate the positive concentrated and negativ e instance separated properties in the graph latent space. It improves the discr iminability by enlarging the positive/negative difference without introducing an y additional computational cost while maintaining high learning efficiency. To a

ddress the limited supervision using data augmentation, a smooth variational rec
onstruction loss is introduced by modeling the intra-instance variance, which im
proves the robustness. Extensive experiments demonstrate the superiority of PSLR
 over state-of-the-art unsupervised methods on both seen and unseen categories w
ith cosine similarity. Code is available at https://github.com/mangye16/PSLR
****

How Useful Is Self-Supervised Pretraining for Visual Tasks?
Alejandro Newell, Jia Deng; Proceedings of the IEEE/CVF Conference on Computer
Vision and Pattern Recognition (CVPR), 2020, pp. 7345-7354
Recent advances have spurred incredible progress in self-supervised pretraining
for vision. We investigate what factors may play a role in the utility of these
pretraining methods for practitioners. To do this, we evaluate various self-supe
rvised algorithms across a comprehensive array of synthetic datasets and downstr
eam tasks. We prepare a suite of synthetic data that enables an endless supply o
f annotated images as well as full control over dataset difficulty. Our experime
nts offer insights into how the utility of self-supervision changes as the numbe
r of available labels grows as well as how the utility changes as a function of
the downstream task and the properties of the training data. We also find that l
inear evaluation does not correlate with finetuning performance. Code and data i
s available at \href https://www.github.com/princeton-vl/selfstudy github.com/p
rinceton-vl/selfstudy .
****

Action Segmentation With Joint Self-Supervised Temporal Domain Adaptation
Min-Hung Chen, Baopu Li, Yingze Bao, Ghassan AlRegib, Zsolt Kira; Proceeding
s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
2020, pp. 9454-9463
Despite the recent progress of fully-supervised action segmentation techniques,
the performance is still not fully satisfactory. One main challenge is the probl
em of spatiotemporal variations (e.g. different people may perform the same acti
vity in various ways). Therefore, we exploit unlabeled videos to address this pr
oblem by reformulating the action segmentation task as a cross-domain problem wi
th domain discrepancy caused by spatio-temporal variations. To reduce the discre
pancy, we propose SelfSupervised Temporal Domain Adaptation (SSTDA), which conta
ins two self-supervised auxiliary tasks (binary and sequential domain prediction
) to jointly align cross-domain feature spaces embedded with local and global te
mporal dynamics, achieving better performance than other Domain Adaptation (DA)
approaches. On three challenging benchmark datasets (GTEA, 50Salads, and Breakfa
st), SSTDA outperforms the current state-of-the-art method by large margins (e.g
. for the F1@25 score, from 59.6% to 69.1% on Breakfast, from 73.4% to 81.5% on
50Salads, and from 83.6% to 89.1% on GTEA), and requires only 65% of the labeled
 training data for comparable performance, demonstrating the usefulness of adapt
ing to unlabeled target videos across variations. The source code is available a
t https://github.com/cmhungsteve/SSTDA.
****

Regularizing Discriminative Capability of CGANs for Semi-Supervised Generative L
earning
Yi Liu, Guangchang Deng, Xiangping Zeng, Si Wu, Zhiwen Yu, Hau-San Wong; Pr
oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
(CVPR), 2020, pp. 5720-5729
Semi-supervised generative learning aims to learn the underlying class-condition
al distribution of partially labeled data. Generative Adversarial Networks (GANs
) have led to promising progress in this task. However, it still needs to furthe
r explore the issue of imbalance between real labeled data and fake data in the
adversarial learning process. To address this issue, we propose a regularization
 technique based on Random Regional Replacement (R^3-regularization) to facilita
te the generative learning process. Specifically, we construct two types of betw
een-class instances: cross-category ones and real-fake ones. These instances cou
ld be closer to the decision boundaries and are important for regularizing the c
lassification and discriminative networks in our class-conditional GANs, which w
e refer to as R^3-CGAN. Better guidance from these two networks makes the genera

tive network produce instances with class-specific information and high fidelity. We experiment with multiple standard benchmarks, and demonstrate that the R^3-regularization can lead to significant improvement in both classification and class-conditional image synthesis.
****

State-Relabeling Adversarial Active Learning
Beichen Zhang, Liang Li, Shijie Yang, Shuhui Wang, Zheng-Jun Zha, Qingming Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8756-8765
Active learning is to design label-efficient algorithms by sampling the most representative samples to be labeled by an oracle. In this paper, we propose a state relabeling adversarial active learning model (SRAAL), that leverages both the annotation and the labeled/unlabeled state information for deriving the most informative unlabeled samples. The SRAAL consists of a representation generator and a state discriminator. The generator uses the complementary annotation information with traditional reconstruction information to generate the unified representation of samples, which embeds the semantic into the whole data representation. Then, we design an online uncertainty indicator in the discriminator, which endues unlabeled samples with different importance. As a result, we can select the most informative samples based on the discriminator's predicted state. We also design an algorithm to initialize the labeled pool, which makes subsequent sampling more efficient. The experiments conducted on various datasets show that our model outperforms the previous state-of-art active learning methods and our initially sampling algorithm achieves better performance.
****

Group Sparsity: The Hinge Between Filter Pruning and Decomposition for Network Compression
Yawei Li, Shuhang Gu, Christoph Mayer, Luc Van Gool, Radu Timofte; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8018-8027
In this paper, we analyze two popular network compression techniques, i.e. filter pruning and low-rank decomposition, in a unified sense. By simply changing the way the sparsity regularization is enforced, filter pruning and low-rank decomposition can be derived accordingly. This provides another flexible choice for network compression because the techniques complement each other. For example, in popular network architectures with shortcut connections (e.g. ResNet), filter pruning cannot deal with the last convolutional layer in a ResBlock while the low-rank decomposition methods can. In addition, we propose to compress the whole network jointly instead of in a layer-wise manner. Our approach proves its potential as it compares favorably to the state-of-the-art on several benchmarks. Code is available at https://github.com/ofsoundof/group_sparsity.
****

P-nets: Deep Polynomial Neural Networks
Grigorios G. Chrysos, Stylianos Moschoglou, Giorgos Bouritsas, Yannis Panagakis, Jiankang Deng, Stefanos Zafeiriou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7325-7335
Deep Convolutional Neural Networks (DCNNs) is currently the method of choice both for generative, as well as for discriminative learning in computer vision and machine learning. The success of DCNNs can be attributed to the careful selection of their building blocks (e.g., residual blocks, rectifiers, sophisticated normalization schemes, to mention but a few). In this paper, we propose \Pi-Nets, a new class of DCNNs. \Pi-Nets are polynomial neural networks, i.e., the output is a high-order polynomial of the input. \Pi-Nets can be implemented using special kind of skip connections and their parameters can be represented via high-order tensors. We empirically demonstrate that \Pi-Nets have better representation power than standard DCNNs and they even produce good results without the use of non-linear activation functions in a large battery of tasks and signals, i.e., images, graphs, and audio. When used in conjunction with activation functions, \Pi-Nets produce state-of-the-art results in challenging tasks, such as image generation. Lastly, our framework elucidates why recent generative models, such as St

yleGAN, improve upon their predecessors, e.g., ProGAN.
****

G3AN: Disentangling Appearance and Motion for Video Generation
Yaohui Wang, Piotr Bilinski, Francois Bremond, Antitza Dantcheva; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5264-5273
Creating realistic human videos entails the challenge of being able to simultaneously generate both appearance, as well as motion. To tackle this challenge, we introduce G3AN, a novel spatio-temporal generative model, which seeks to capture the distribution of high dimensional video data and to model appearance and motion in disentangled manner. The latter is achieved by decomposing appearance and motion in a three-stream Generator, where the main stream aims to model spatio-temporal consistency, whereas the two auxiliary streams augment the main stream with multi-scale appearance and motion features, respectively. An extensive quantitative and qualitative analysis shows that our model systematically and significantly outperforms state-of-the-art methods on the facial expression datasets MUG and UvA-NEMO, as well as the Weizmann and UCF101 datasets on human action. Additional analysis on the learned latent representations confirms the successful decomposition of appearance and motion.
****

DeepFaceFlow: In-the-Wild Dense 3D Facial Motion Estimation
Mohammad Rami Koujan, Anastasios Roussos, Stefanos Zafeiriou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6618-6627
Dense 3D facial motion capture from only monocular in-the-wild pairs of RGB images is a highly challenging problem with numerous applications, ranging from facial expression recognition to facial reenactment. In this work, we propose DeepFaceFlow, a robust, fast, and highly-accurate framework for the dense estimation of 3D non-rigid facial flow between pairs of monocular images. Our DeepFaceFlow framework was trained and tested on two very large-scale facial video datasets, one of them of our own collection and annotation, with the aid of occlusion-aware and 3D-based loss function. We conduct comprehensive experiments probing different aspects of our approach and demonstrating its improved performance against state-of-the-art flow and 3D reconstruction methods. Furthermore, we incorporate our framework in a full-head state-of-the-art facial video synthesis method and demonstrate the ability of our method in better representing and capturing the facial dynamics, resulting in a highly-realistic facial video synthesis. Given registered pairs of images, our framework generates 3D flow maps at 60 fps.
****

TubeTK: Adopting Tubes to Track Multi-Object in a One-Step Training Model
Bo Pang, Yizhuo Li, Yifan Zhang, Muchen Li, Cewu Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6308-6318
Multi-object tracking is a fundamental vision problem that has been studied for a long time. As deep learning brings excellent performances to object detection algorithms, Tracking by Detection (TBD) has become the mainstream tracking framework. Despite the success of TBD, this two-step method is too complicated to train in an end-to-end manner and induces many challenges as well, such as insufficient exploration of video spatial-temporal information, vulnerability when facing object occlusion, and excessive reliance on detection results. To address these challenges, we propose a concise end-to-end model TubeTK which only needs one step training by introducing the "bounding-tube" to indicate temporal-spatial locations of objects in a short video clip. TubeTK provides a novel direction of multi-object tracking, and we demonstrate its potential to solve the above challenges without bells and whistles. We analyze the performance of TubeTK on several MOT benchmarks and provide empirical evidence to show that TubeTK has the ability to overcome occlusions to some extent without any ancillary technologies like Re-ID. Compared with other methods that adopt private detection results, our one-stage end-to-end model achieves state-of-the-art performances even if it adopts no ready-made detection results. We hope that the proposed TubeTK model can se

rve as a simple but strong alternative for video-based MOT task. The code and mo
del will be publicly available accompanying this paper.
****

Few-Shot Open-Set Recognition Using Meta-Learning
Bo Liu, Hao Kang, Haoxiang Li, Gang Hua, Nuno Vasconcelos; Proceedings of th
e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, p
p. 8798-8807
The problem of open-set recognition is considered. While previous approaches onl
y consider this problem in the context of large-scale classifier training, we se
ek a unified solution for this and the low-shot classification setting. It is ar
gued that the classic softmax classifier is a poor solution for open-set recogni
tion, since it tends to overfit on the training classes. Randomization is then p
roposed as a solution to this problem. This suggests the use of meta-learning te
chniques, commonly used for few-shot classification, for the solution of open-se
t recognition. A new oPen sEt mEta LEaRning (PEELER) algorithm is then introduce
d. This combines the random selection of a set of novel classes per episode, a l
oss that maximizes the posterior entropy for examples of those classes, and a ne
w metric learning formulation based on the Mahalanobis distance. Experimental re
sults show that PEELER achieves state of the art open set recognition performanc
e for both few-shot and large-scale recognition. On CIFAR and miniImageNet, it a
chieves substantial gains in seen/unseen class detection AUROC for a given seen-
class classification accuracy.
****

Sequential 3D Human Pose and Shape Estimation From Point Clouds
Kangkan Wang, Jin Xie, Guofeng Zhang, Lei Liu, Jian Yang; Proceedings of the
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp
. 7275-7284
This work addresses the problem of 3D human pose and shape estimation from a seq
uence of point clouds. Existing sequential 3D human shape estimation methods mai
nly focus on the template model fitting from a sequence of depth images or the p
arametric model regression from a sequence of RGB images. In this paper, we prop
ose a novel sequential 3D human pose and shape estimation framework from a seque
nce of point clouds. Specifically, the proposed framework can regress 3D coordin
ates of mesh vertices at different resolutions from the latent features of point
clouds. Based on the estimated 3D coordinates and features at the low resolutio
n, we develop a spatial-temporal mesh attention convolution (MAC) to predict the
3D coordinates of mesh vertices at the high resolution. By assigning specific a
ttentional weights to different neighboring points in the spatial and temporal d
omains, our spatial-temporal MAC can capture structured spatial and temporal fea
tures of point clouds. We further generalize our framework to the real data of h
uman bodies with a weakly supervised fine-tuning method. The experimental result
s on SURREAL, Human3.6M, DFAUST and the real detailed data demonstrate that the
proposed approach can accurately recover the 3D body model sequence from a seque
nce of point clouds.
****

Sequential Mastery of Multiple Visual Tasks: Networks Naturally Learn to Learn a
nd Forget to Forget
Guy Davidson, Michael C. Mozer; Proceedings of the IEEE/CVF Conference on Compu
ter Vision and Pattern Recognition (CVPR), 2020, pp. 9282-9293
We explore the behavior of a standard convolutional neural net in a continual-le
arning setting that introduces visual classification tasks sequentially and requ
ires the net to master new tasks while preserving mastery of previously learned
tasks. This setting corresponds to that which human learners face as they acquir
e domain expertise serially, for example, as an individual studies a textbook. T
hrough simulations involving sequences of ten related visual tasks, we find reas
on for optimism that nets will scale well as they advance from having a single s
kill to becoming multi-skill domain experts. We observe two key phenomena. First
, forward facilitation---the accelerated learning of task n+1 having learned n p
revious tasks---grows with n. Second, backward interference---the forgetting of
the n previous tasks when learning task n+1 ---diminishes with n. Amplifying for

ward facilitation is the goal of research on metalearning, and attenuating backward interference is the goal of research on catastrophic forgetting. We find that both of these goals are attained simply through broader exposure to a domain.
****

Siam R-CNN: Visual Tracking by Re-Detection
Paul Voigtlaender, Jonathon Luiten, Philip H.S. Torr, Bastian Leibe; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6578-6588
We present Siam R-CNN, a Siamese re-detection architecture which unleashes the full power of two-stage object detection approaches for visual object tracking. We combine this with a novel tracklet-based dynamic programming algorithm, which takes advantage of re-detections of both the first-frame template and previous-frame predictions, to model the full history of both the object to be tracked and potential distractor objects. This enables our approach to make better tracking decisions, as well as to re-detect tracked objects after long occlusion. Finally, we propose a novel hard example mining strategy to improve Siam R-CNN's robustness to similar looking objects. Siam R-CNN achieves the current best performance on ten tracking benchmarks, with especially strong results for long-term tracking. We make our code and models available at www.vision.rwth-aachen.de/page/siamrcnn.
****

Eternal Sunshine of the Spotless Net: Selective Forgetting in Deep Networks
Aditya Golatkar, Alessandro Achille, Stefano Soatto; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9304-9312
We explore the problem of selectively forgetting a particular subset of the data used for training a deep neural network. While the effects of the data to be forgotten can be hidden from the output of the network, insights may still be gleaned by probing deep into its weights. We propose a method for "scrubbing" the weights clean of information about a particular set of training data. The method does not require retraining from scratch, nor access to the data originally used for training. Instead, the weights are modified so that any probing function of the weights is indistinguishable from the same function applied to the weights of a network trained without the data to be forgotten. This condition is a generalized and weaker form of Differential Privacy. Exploiting ideas related to the stability of stochastic gradient descent, we introduce an upper-bound on the amount of information remaining in the weights, which can be estimated efficiently even for deep neural networks.
****

MSG-GAN: Multi-Scale Gradients for Generative Adversarial Networks
Animesh Karnewar, Oliver Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7799-7808
While Generative Adversarial Networks (GANs) have seen huge successes in image synthesis tasks, they are notoriously difficult to adapt to different datasets, in part due to instability during training and sensitivity to hyperparameters. One commonly accepted reason for this instability is that gradients passing from the discriminator to the generator become uninformative when there isn't enough overlap in the supports of the real and fake distributions. In this work, we propose the Multi-Scale Gradient Generative Adversarial Network (MSG-GAN), a simple but effective technique for addressing this by allowing the flow of gradients from the discriminator to the generator at multiple scales. This technique provides a stable approach for high resolution image synthesis, and serves as an alternative to the commonly used progressive growing technique. We show that MSG-GAN converges stably on a variety of image datasets of different sizes, resolutions and domains, as well as different types of loss functions and architectures, all with the same set of fixed hyperparameters. When compared to state-of-the-art GANs, our approach matches or exceeds the performance in most of the cases we tried.
****

Transferring Cross-Domain Knowledge for Video Sign Language Recognition

Dongxu Li, Xin Yu, Chenchen Xu, Lars Petersson, Hongdong Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6205-6214

Word-level sign language recognition (WSLR) is a fundamental task in sign language interpretation. It requires models to recognize isolated sign words from videos. However, annotating WSLR data needs expert knowledge, thus limiting WSLR dataset acquisition. On the contrary, there are abundant subtitled sign news videos on the internet. Since these videos have no word-level annotation and exhibit a large domain gap from isolated signs, they cannot be directly used for training WSLR models. We observe that despite the existence of a large domain gap, isolated and news signs share the same visual concepts, such as hand gestures and body movements. Motivated by this observation, we propose a novel method that learns domain-invariant visual concepts and fertilizes WSLR models by transferring knowledge of subtitled news sign to them. To this end, we extract news signs using a base WSLR model, and then design a classifier jointly trained on news and isolated signs to coarsely align these two domain features. In order to learn domain-invariant features within each class and suppress domain-specific features, our method further resorts to an external memory to store the class centroids of the aligned news signs. We then design a temporal attention based on the learnt descriptor to improve recognition performance. Experimental results on standard WSLR datasets show that our method outperforms previous state-of-the-art methods significantly. We also demonstrate the effectiveness of our method on automatically localizing signs from sign news, achieving 28.1 for AP@0.5.

****

Flow Contrastive Estimation of Energy-Based Models

Ruiqi Gao, Erik Nijkamp, Diederik P. Kingma, Zhen Xu, Andrew M. Dai, Ying Nian Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7518-7528

This paper studies a training method to jointly estimate an energy-based model and a flow-based model, in which the two models are iteratively updated based on a shared adversarial value function. This joint training method has the following traits. (1) The update of the energy-based model is based on noise contrastive estimation, with the flow model serving as a strong noise distribution. (2) The update of the flow model approximately minimizes the Jensen-Shannon divergence between the flow model and the data distribution. (3) Unlike generative adversarial networks (GAN) which estimates an implicit probability distribution defined by a generator model, our method estimates two explicit probabilistic distributions on the data. Using the proposed method we demonstrate a significant improvement on the synthesis quality of the flow model, and show the effectiveness of unsupervised feature learning by the learned energy-based model. Furthermore, the proposed training method can be easily adapted to semi-supervised learning. We achieve competitive results to the state-of-the-art semi-supervised learning methods.

****

Improving the Robustness of Capsule Networks to Image Affine Transformations

Jindong Gu, Volker Tresp; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7285-7293

Convolutional neural networks (CNNs) achieve translational invariance by using pooling operations. However, the operations do not preserve the spatial relationships in the learned representations. Hence, CNNs cannot extrapolate to various geometric transformations of inputs. Recently, Capsule Networks (CapsNets) have been proposed to tackle this problem. In CapsNets, each entity is represented by a vector and routed to high-level entity representations by a dynamic routing algorithm. CapsNets have been shown to be more robust than CNNs to affine transformations of inputs. However, there is still a huge gap between their performance on transformed inputs compared to untransformed versions. In this work, we first revisit the routing procedure by (un)rolling its forward and backward passes. Our investigation reveals that the routing procedure contributes neither to the generalization ability nor to the affine robustness of the CapsNets. Furthermore, we explore the limitations of capsule transformations and propose affine CapsNe

ts (Aff-CapsNets), which are more robust to affine transformations. On our bench mark task, where models are trained on the MNIST dataset and tested on the AffNI ST dataset, our Aff-CapsNets improve the benchmark performance by a large margin (from 79% to 93.21%), without using any routing mechanism.
****

Interactive Two-Stream Decoder for Accurate and Fast Saliency Detection
Huajun Zhou, Xiaohua Xie, Jian-Huang Lai, Zixuan Chen, Lingxiao Yang; Procee dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP R), 2020, pp. 9141-9150
Recently, contour information largely improves the performance of saliency detec tion. However, the discussion on the correlation between saliency and contour re mains scarce. In this paper, we first analyze such correlation and then propose an interactive two-stream decoder to explore multiple cues, including saliency, contour and their correlation. Specifically, our decoder consists of two branche s, a saliency branch and a contour branch. Each branch is assigned to learn dist inctive features for predicting the corresponding map. Meanwhile, the intermedia te connections are forced to learn the correlation by interactively transmitting the features from each branch to the other one. In addition, we develop an adap tive contour loss to automatically discriminate hard examples during learning pr ocess. Extensive experiments on six benchmarks well demonstrate that our network achieves competitive performance with a fast speed around 50 FPS. Moreover, our VGG-based model only contains 17.08 million parameters, which is significantly smaller than other VGG-based approaches. Code has been made available at: https: //github.com/moothes/ITSD-pytorch.
****

ViewAL: Active Learning With Viewpoint Entropy for Semantic Segmentation
Yawar Siddiqui, Julien Valentin, Matthias Niessner; Proceedings of the IEEE/CV F Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9433-9 443
We propose ViewAL, a novel active learning strategy for semantic segmentation th at exploits viewpoint consistency in multi-view datasets. Our core idea is that inconsistencies in model predictions across viewpoints provide a very reliable m easure of uncertainty and encourage the model to perform well irrespective of th e viewpoint under which objects are observed. To incorporate this uncertainty me asure, we introduce a new viewpoint entropy formulation, which is the basis of o ur active learning strategy. In addition, we propose uncertainty computations on a superpixel level, which exploits inherently localized signal in the segmentat ion task, directly lowering the annotation costs. This combination of viewpoint entropy and the use of superpixels allows to efficiently select samples that are highly informative for improving the network. We demonstrate that our proposed active learning strategy not only yields the best-performing models for the same amount of required labeled data, but also significantly reduces labeling effort . For instance, our method achieves 95% of maximum achievable network performanc e using only 7%, 17%, and 24% labeled data on SceneNet-RGBD, ScanNet, and Matter port3D, respectively. On these datasets, the best state-of-the-art method achiev es the same performance with 14%, 27% and 33% labeled data. Finally, we demonstr ate that labeling using superpixels yields the same quality of ground-truth comp ared to labeling whole images, but requires 25% less time.
****

A U-Net Based Discriminator for Generative Adversarial Networks
Edgar Schonfeld, Bernt Schiele, Anna Khoreva; Proceedings of the IEEE/CVF Conf erence on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8207-8216
Among the major remaining challenges for generative adversarial networks (GANs) is the capacity to synthesize globally and locally coherent images with object s hapes and textures indistinguishable from real images. To target this issue we p ropose an alternative U-Net based discriminator architecture, borrowing the insi ghts from the segmentation literature. The proposed U-Net based architecture all ows to provide detailed per-pixel feedback to the generator while maintaining th e global coherence of synthesized images, by providing the global image feedback as well. Empowered by the per-pixel response of the discriminator, we further p

ropose a per-pixel consistency regularization technique based on the CutMix data augmentation, encouraging the U-Net discriminator to focus more on semantic and structural changes between real and fake images. This improves the U-Net discriminator training, further enhancing the quality of generated samples. The novel discriminator improves over the state of the art in terms of the standard distribution and image quality metrics, enabling the generator to synthesize images with varying structure, appearance and levels of detail, maintaining global and local realism. Compared to the BigGAN baseline, we achieve an average improvement of 2.7 FID points across FFHQ, CelebA, and the proposed COCO-Animals dataset.
****

Diversified Arbitrary Style Transfer via Deep Feature Perturbation
Zhizhong Wang, Lei Zhao, Haibo Chen, Lihong Qiu, Qihang Mo, Sihuan Lin, Wei Xing, Dongming Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7789-7798
Image style transfer is an underdetermined problem, where a large number of solutions can satisfy the same constraint (the content and style). Although there have been some efforts to improve the diversity of style transfer by introducing an alternative diversity loss, they have restricted generalization, limited diversity and poor scalability. In this paper, we tackle these limitations and propose a simple yet effective method for diversified arbitrary style transfer. The key idea of our method is an operation called deep feature perturbation (DFP), which uses an orthogonal random noise matrix to perturb the deep image feature maps while keeping the original style information unchanged. Our DFP operation can be easily integrated into many existing WCT (whitening and coloring transform)-based methods, and empower them to generate diverse results for arbitrary styles. Experimental results demonstrate that this learning-free and universal method can greatly increase the diversity while maintaining the quality of stylization.
****

15 Keypoints Is All You Need
Michael Snower, Asim Kadav, Farley Lai, Hans Peter Graf; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6738-6748
Pose-tracking is an important problem that requires identifying unique human pose-instances and matching them temporally across different frames in a video. However, existing pose-tracking methods are unable to accurately model temporal relationships and require significant computation, often computing the tracks offline. We present an efficient multi-person pose-tracking method, KeyTrack that only relies on keypoint information without using any RGB or optical flow to locate and track human keypoints in real-time. KeyTrack is a top-down approach that learns spatio-temporal pose relationships by modeling the multi-person pose-tracking problem as a novel Pose Entailment task using a Transformer based architecture. Furthermore, KeyTrack uses a novel, parameter-free, keypoint refinement technique that improves the keypoint estimates used by the Transformers. We achieve state-of-the-art results on PoseTrack'17 and PoseTrack'18 benchmarks while using only a fraction of the computation used by most other methods for computing the tracking information.
****

LUVLi Face Alignment: Estimating Landmarks' Location, Uncertainty, and Visibility Likelihood
Abhinav Kumar, Tim K. Marks, Wenxuan Mou, Ye Wang, Michael Jones, Anoop Cherian, Toshiaki Koike-Akino, Xiaoming Liu, Chen Feng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8236-8246
Modern face alignment methods have become quite accurate at predicting the locations of facial landmarks, but they do not typically estimate the uncertainty of their predicted locations nor predict whether landmarks are visible. In this paper, we present a novel framework for jointly predicting landmark locations, associated uncertainties of these predicted locations, and landmark visibilities. We model these as mixed random variables and estimate them using a deep network trained using our proposed Location, Uncertainty, and Visibility Likelihood (LUVLi

) loss. In addition, we release an entirely new labeling of a large face alignme nt dataset with over 19,000 face images in a full range of head poses. Each face is manually labeled with the ground-truth locations of 68 landmarks, with the a dditional information of whether each landmarks is visible, self-occluded (due t o extreme head poses), or externally occluded. Not only does our joint estimatio n yield accurate estimates of the uncertainty of predicted landmark locations, b ut it also yields state-of-the-art estimates for the landmark locations themselv es on mulitple standard face alignment datasets. Our method's estimates of the u ncertainty of predicted landmark locations could be used to automatically identi fy input images on which face alignment fails, which can be critical for downstr eam tasks.
****

Learning to Cartoonize Using White-Box Cartoon Representations
Xinrui Wang, Jinze Yu; Proceedings of the IEEE/CVF Conference on Computer Visio n and Pattern Recognition (CVPR), 2020, pp. 8090-8099
This paper presents an approach for image cartoonization. By observing the carto on painting behavior and consulting artists, we propose to separately identify t hree white-box representations from images: the surface representation that cont ains smooth surface of cartoon images, the structure representation that refers to the sparse color-blocks and flatten global content in the celluloid style wor kflow, and the texture representation that reflects high-frequency texture, cont ours and details in cartoon images. A Generative Adversarial Network (GAN) frame work is used to learn the extracted representations and to cartoonize images. Th e learning objectives of our method are separately based on each extracted repre sentations, making our framework controllable and adjustable. This enables our a pproach to meet artists' requirements in different styles and diverse use cases. Qualitative comparisons and quantitative analyses, as well as user studies, hav e been conducted to validate the effectiveness of this approach, and our method outperforms previous methods in all comparisons. Finally, the ablation study dem onstrates the influence of each component in our framework.
****

PointAugment: An Auto-Augmentation Framework for Point Cloud Classification
Ruihui Li, Xianzhi Li, Pheng-Ann Heng, Chi-Wing Fu; Proceedings of the IEEE/C VF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6378-6387
We present PointAugment, a new auto-augmentation framework that automatically op timizes and augments point cloud samples to enrich the data diversity when we tr ain a classification network. Different from existing auto-augmentation methods for 2D images, PointAugment is sample-aware and takes an adversarial learning st rategy to jointly optimize an augmentor network and a classifier network, such t hat the augmentor can learn to produce augmented samples that best fit the class ifier. Moreover, we formulate a learnable point augmentation function with a sha pe-wise transformation and a point-wise displacement, and carefully design loss functions to adopt the augmented samples based on the learning progress of the c lassifier. Extensive experiments also confirm PointAugment's effectiveness and r obustness to improve the performance of various networks on shape classification and retrival.
****

Siamese Box Adaptive Network for Visual Tracking
Zedu Chen, Bineng Zhong, Guorong Li, Shengping Zhang, Rongrong Ji; Proceedin gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6668-6677
Most of the existing trackers usually rely on either a multi-scale searching sch eme or pre-defined anchor boxes to accurately estimate the scale and aspect rati o of a target. Unfortunately, they typically call for tedious and heuristic conf igurations. To address this issue, we propose a simple yet effective visual trac king framework (named Siamese Box Adaptive Network, SiamBAN) by exploiting the e xpressive power of the fully convolutional network (FCN). SiamBAN views the visu al tracking problem as a parallel classification and regression problem, and thu s directly classifies objects and regresses their bounding boxes in a unified FC

N. The no-prior box design avoids hyper-parameters associated with the candidate boxes, making SiamBAN more flexible and general. Extensive experiments on visual tracking benchmarks including VOT2018, VOT2019, OTB100, NFS, UAV123, and LaSOT demonstrate that SiamBAN achieves state-of-the-art performance and runs at 40 FPS, confirming its effectiveness and efficiency. The code will be available at https://github.com/hqucv/siamban.
****

Interpretable and Accurate Fine-grained Recognition via Region Grouping
Zixuan Huang, Yin Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8662-8672
We present an interpretable deep model for fine-grained visual recognition. At the core of our method lies the integration of region-based part discovery and attribution within a deep neural network. Our model is trained using image-level object labels, and provides an interpretation of its results via the segmentation of object parts and the identification of their contributions towards classification. To facilitate the learning of object parts without direct supervision, we explore a simple prior of the occurrence of object parts. We demonstrate that this prior, when combined with our region-based part discovery and attribution, leads to an interpretable model that remains highly accurate. Our model is evaluated on major fine-grained recognition datasets, including CUB-200, CelebA and iNaturalist. Our results compares favourably to state-of-the-art methods on classification tasks, and outperforms previous approaches on the localization of object parts.
****

Low-Rank Compression of Neural Nets: Learning the Rank of Each Layer
Yerlan Idelbayev, Miguel A. Carreira-Perpinan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8049-8059
Neural net compression can be achieved by approximating each layer's weight matrix by a low-rank matrix. The real difficulty in doing this is not in training the resulting neural net (made up of one low-rank matrix per layer), but in determining what the optimal rank of each layer is--effectively, an architecture search problem with one hyperparameter per layer. We show that, with a suitable formulation, this problem is amenable to a mixed discrete-continuous optimization jointly over the ranks and over the matrix elements, and give a corresponding algorithm. We show that this indeed can select ranks much better than existing approaches, making low-rank compression much more attractive than previously thought. For example, we can make a VGG network faster than a ResNet and with nearly the same classification error.
****

There and Back Again: Revisiting Backpropagation Saliency Methods
Sylvestre-Alvise Rebuffi, Ruth Fong, Xu Ji, Andrea Vedaldi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8839-8848
Saliency methods seek to explain the predictions of a model by producing an importance map across each input sample. A popular class of such methods is based on backpropagating a signal and analyzing the resulting gradient. Despite much research on such methods, relatively little work has been done to clarify the differences between such methods as well as the desiderata of these techniques. Thus, there is a need for rigorously understanding the relationships between different methods as well as their failure modes. In this work, we conduct a thorough analysis of backpropagation-based saliency methods and propose a single framework under which several such methods can be unified. As a result of our study, we make three additional contributions. First, we use our framework to propose NormGrad, a novel saliency method based on the spatial contribution of gradients of convolutional weights. Second, we combine saliency maps at different layers to test the ability of saliency methods to extract complementary information at different network levels (e.g. trading off spatial resolution and distinctiveness) and we explain why some methods fail at specific layers (e.g., Grad-CAM anywhere besides the last convolutional layer). Third, we introduce a class-sensitivity metric and a meta-learning inspired paradigm applicable to any saliency method for

improving sensitivity to the output class being explained.
****
Learning Meta Face Recognition in Unseen Domains
Jianzhu Guo, Xiangyu Zhu, Chenxu Zhao, Dong Cao, Zhen Lei, Stan Z. Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6163-6172
Face recognition systems are usually faced with unseen domains in real-world applications and show unsatisfactory performance due to their poor generalization. For example, a well-trained model on webface data cannot deal with the ID vs. Spot task in surveillance scenario. In this paper, we aim to learn a generalized model that can directly handle new unseen domains without any model updating. To this end, we propose a novel face recognition method via meta-learning named Meta Face Recognition (MFR). MFR synthesizes the source/target domain shift with a meta-optimization objective, which requires the model to learn effective representations not only on synthesized source domains but also on synthesized target domains. Specifically, we build domain-shift batches through a domain-level sampling strategy and get back-propagated gradients/meta-gradients on synthesized source/target domains by optimizing multi-domain distributions. The gradients and meta-gradients are further combined to update the model to improve generalization. Besides, we propose two benchmarks for generalized face recognition evaluation. Experiments on our benchmarks validate the generalization of our method compared to several baselines and other state-of-the-arts. The proposed benchmarks and code will be available at https://github.com/cleardusk/MFR.
****
MineGAN: Effective Knowledge Transfer From GANs to Target Domains With Few Images
Yaxing Wang, Abel Gonzalez-Garcia, David Berga, Luis Herranz, Fahad Shahbaz Khan, Joost van de Weijer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9332-9341
One of the attractive characteristics of deep neural networks is their ability to transfer knowledge obtained in one domain to other related domains. As a result, high-quality networks can be trained in domains with relatively little training data. This property has been extensively studied for discriminative networks but has received significantly less attention for generative models. Given the often enormous effort required to train GANs, both computationally as well as in the dataset collection, the re-use of pretrained GANs is a desirable objective. We propose a novel knowledge transfer method for generative models based on mining the knowledge that is most beneficial to a specific target domain, either from a single or multiple pretrained GANs. This is done using a miner network that identifies which part of the generative distribution of each pretrained GAN outputs samples closest to the target domain. Mining effectively steers GAN sampling towards suitable regions of the latent space, which facilitates the posterior finetuning and avoids pathologies of other methods such as mode collapse and lack of flexibility. We perform experiments on several complex datasets using various GAN architectures (BigGAN, Progressive GAN) and show that the proposed method, called MineGAN, effectively transfers knowledge to domains with few target images, outperforming existing methods. In addition, MineGAN can successfully transfer knowledge from multiple pretrained GANs. Our code is available at: https://github.com/yaxingwang/MineGAN.
****
State-Aware Tracker for Real-Time Video Object Segmentation
Xi Chen, Zuoxin Li, Ye Yuan, Gang Yu, Jianxin Shen, Donglian Qi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9384-9393
In this work, we address the task of semi-supervised video object segmentation (VOS) and explore how to make efficient use of video property to tackle the challenge of semi-supervision. We propose a novel pipeline called State-Aware Tracker (SAT), which can produce accurate segmentation results with real-time speed. For higher efficiency, SAT takes advantage of the inter-frame consistency and deals with each target object as a tracklet. For more stable and robust performance

over video sequences, SAT gets awareness for each state and makes self-adaptatio n via two feedback loops. One loop assists SAT in generating more stable trackle ts. The other loop helps to construct a more robust and holistic target represen tation. SAT achieves a promising result of 72.3% J&F mean with 39 FPS on DAVIS 2 017-Val dataset, which shows a decent trade-off between efficiency and accuracy.
****

DualSDF: Semantic Shape Manipulation Using a Two-Level Representation
Zekun Hao, Hadar Averbuch-Elor, Noah Snavely, Serge Belongie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7631-7641
We are seeing a Cambrian explosion of 3D shape representations for use in machin e learning. Some representations seek high expressive power in capturing high-re solution detail. Other approaches seek to represent shapes as compositions of si mple parts, which are intuitive for people to understand and easy to edit and ma nipulate. However, it is difficult to achieve both fidelity and interpretability in the same representation. We propose DualSDF, a representation expressing sha pes at two levels of granularity, one capturing fine details and the other repre senting an abstracted proxy shape using simple and semantically consistent shape primitives. To achieve a tight coupling between the two representations, we use a variational objective over a shared latent space. Our two-level model gives r ise to a new shape manipulation technique in which a user can interactively mani pulate the coarse proxy shape and see the changes instantly mirrored in the high -resolution shape. Moreover, our model actively augments and guides the manipula tion towards producing semantically meaningful shapes, making complex manipulati ons possible with minimal user input.
****

Can We Learn Heuristics for Graphical Model Inference Using Reinforcement Learni ng?
Safa Messaoud, Maghav Kumar, Alexander G. Schwing; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7589-75 99
Combinatorial optimization is frequently used in computer vision. For instance, in applications like semantic segmentation, human pose estimation and action rec ognition, programs are formulated for solving inference in Conditional Random Fi elds (CRFs) to produce a structured output that is consistent with visual featur es of the image. However, solving inference in CRFs is in general intractable, a nd approximation methods are computationally demanding and limited to unary, pai rwise and hand-crafted forms of higher order potentials. In this paper, we show that we can learn program heuristics, i.e., policies, for solving inference in h igher order CRFs for the task of semantic segmentation, using reinforcement lear ning. Our method solves inference tasks efficiently without imposing any constra ints on the form of the potentials. We show compelling results on the Pascal VOC and MOTS datasets.
****

D3S - A Discriminative Single Shot Segmentation Tracker
Alan Lukezic, Jiri Matas, Matej Kristan; Proceedings of the IEEE/CVF Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7133-7142
Template-based discriminative trackers are currently the dominant tracking parad igm due to their robustness, but are restricted to bounding box tracking and a l imited range of transformation models, which reduces their localization accuracy . We propose a discriminative single-shot segmentation tracker - D3S, which narr ows the gap between visual object tracking and video object segmentation. A sing le-shot network applies two target models with complementary geometric propertie s, one invariant to a broad range of transformations, including non-rigid deform ations, the other assuming a rigid object to simultaneously achieve high robustn ess and online target segmentation. Without per-dataset finetuning and trained o nly for segmentation as the primary output, D3S outperforms all trackers on VOT2 016, VOT2018 and GOT-10k benchmarks and performs close to the state-of-the-art t rackers on the TrackingNet. D3S outperforms the leading segmentation tracker Sia mMask on video segmentation benchmark and performs on par with top video object

segmentation algorithms, while running an order of magnitude faster, close to re
al-time.
****

Cross-Spectral Face Hallucination via Disentangling Independent Factors
Boyan Duan, Chaoyou Fu, Yi Li, Xingguang Song, Ran He; Proceedings of the IE
EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7
930-7938
The cross-sensor gap is one of the challenges that have aroused much research in
terests in Heterogeneous Face Recognition (HFR). Although recent methods have at
tempted to fill the gap with deep generative networks, most of them suffer from
the inevitable misalignment between different face modalities. Instead of imagin
g sensors, the misalignment primarily results from facial geometric variations t
hat are independent of the spectrum. Rather than building a monolithic but compl
ex structure, this paper proposes a Pose Aligned Cross-spectral Hallucination (P
ACH) approach to disentangle the independent factors and deal with them in indiv
idual stages. In the first stage, an Unsupervised Face Alignment (UFA) module is
designed to align the facial shapes of the near-infrared (NIR) images with thos
e of the visible (VIS) images in a generative way, where UV maps are effectively
utilized as the shape guidance. Thus the task of the second stage becomes spect
rum translation with aligned paired data. We develop a Texture Prior Synthesis (
TPS) module to achieve complexion control and consequently generate more realist
ic VIS images than existing methods. Experiments on three challenging NIR-VIS da
tasets verify the effectiveness of our approach in producing visually appealing
images and achieving state-of-the-art performance in HFR.
****

Deep Face Super-Resolution With Iterative Collaboration Between Attentive Recove
ry and Landmark Estimation
Cheng Ma, Zhenyu Jiang, Yongming Rao, Jiwen Lu, Jie Zhou; Proceedings of the
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp
. 5569-5578
Recent works based on deep learning and facial priors have succeeded in super-re
solving severely degraded facial images. However, the prior knowledge is not ful
ly exploited in existing methods, since facial priors such as landmark and compo
nent maps are always estimated by low-resolution or coarsely super-resolved imag
es, which may be inaccurate and thus affect the recovery performance. In this pa
per, we propose a deep face super-resolution (FSR) method with iterative collabo
ration between two recurrent networks which focus on facial image recovery and l
andmark estimation respectively. In each recurrent step, the recovery branch uti
lizes the prior knowledge of landmarks to yield higher-quality images which faci
litate more accurate landmark estimation in turn. Therefore, the iterative infor
mation interaction between two processes boosts the performance of each other pr
ogressively. Moreover, a new attentive fusion module is designed to strengthen t
he guidance of landmark maps, where facial components are generated individually
and aggregated attentively for better restoration. Quantitative and qualitative
experimental results show the proposed method significantly outperforms state-o
f-the-art FSR methods in recovering high-quality face images.
****

Weakly-Supervised 3D Human Pose Learning via Multi-View Images in the Wild
Umar Iqbal, Pavlo Molchanov, Jan Kautz; Proceedings of the IEEE/CVF Conference
on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5243-5252
One major challenge for monocular 3D human pose estimation in-the-wild is the ac
quisition of training data that contains unconstrained images annotated with acc
urate 3D poses. In this paper, we address this challenge by proposing a weakly-s
upervised approach that does not require 3D annotations and learns to estimate 3
D poses from unlabeled multi-view data, which can be acquired easily in in-the-w
ild environments. We propose a novel end-to-end learning framework that enables
weakly-supervised training using multi-view consistency. Since multi-view consis
tency is prone to degenerated solutions, we adopt a 2.5D pose representation and
propose a novel objective function that can only be minimized when the predicti
ons of the trained model are consistent and plausible across all camera views. W

e evaluate our proposed approach on two large scale datasets (Human3.6M and MPII -INF-3DHP) where it achieves state-of-the-art performance among semi-/weakly-sup ervised methods.
****

Data Uncertainty Learning in Face Recognition
Jie Chang, Zhonghao Lan, Changmao Cheng, Yichen Wei; Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5710 -5719
Modeling data uncertainty is important for noisy images, but seldom explored for face recognition. The pioneer work, PFE, considers uncertainty by modeling each face image embedding as a Gaussian distribution. It is quite effective. However , it uses fixed feature (mean of the Gaussian) from an existing model. It only e stimates the variance and relies on an ad-hoc and costly metric. Thus, it is not easy to use. It is unclear how uncertainty affects feature learning. This work applies data uncertainty learning to face recognition, such that the feature (me an) and uncertainty (variance) are learnt simultaneously, for the first time. Tw o learning methods are proposed. They are easy to use and outperform existing de terministic methods as well as PFE on challenging unconstrained scenarios. We al so provide insightful analysis on how incorporating uncertainty estimation helps reducing the adverse effects of noisy samples and affects the feature learning.
****

Learning Fast and Robust Target Models for Video Object Segmentation
Andreas Robinson, Felix Jaremo Lawin, Martin Danelljan, Fahad Shahbaz Khan, Michael Felsberg; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7406-7415
Video object segmentation (VOS) is a highly challenging problem since the initia l mask, defining the target object, is only given at test-time. The main difficu lty is to effectively handle appearance changes and similar background objects, while maintaining accurate segmentation. Most previous approaches fine-tune segm entation networks on the first frame, resulting in impractical frame-rates and r isk of overfitting. More recent methods integrate generative target appearance m odels, but either achieve limited robustness or require large amounts of trainin g data. We propose a novel VOS architecture consisting of two network components . The target appearance model consists of a light-weight module, which is learne d during the inference stage using fast optimization techniques to predict a coa rse but robust target segmentation. The segmentation model is exclusively traine d offline, designed to process the coarse scores into high quality segmentation masks. Our method is fast, easily trainable and remains highly effective in case s of limited training data. We perform extensive experiments on the challenging YouTube-VOS and DAVIS datasets. Our network achieves favorable performance, whil e operating at higher frame-rates compared to state-of-the-art. Code and trained models are available at https://github.com/andr345/frtm-vos.
****

Transferring and Regularizing Prediction for Semantic Segmentation
Yiheng Zhang, Zhaofan Qiu, Ting Yao, Chong-Wah Ngo, Dong Liu, Tao Mei; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2020, pp. 9621-9630
Semantic segmentation often requires a large set of images with pixel-level anno tations. In the view of extremely expensive expert labeling, recent research has shown that the models trained on photo-realistic synthetic data (e.g., computer games) with computer-generated annotations can be adapted to real images. Despi te this progress, without constraining the prediction on real images, the models will easily overfit on synthetic data due to severe domain mismatch. In this pa per, we novelly exploit the intrinsic properties of semantic segmentation to all eviate such problem for model transfer. Specifically, we present a Regularizer o f Prediction Transfer (RPT) that imposes the intrinsic properties as constraints to regularize model transfer in an unsupervised fashion. These constraints incl ude patch-level, cluster-level and context-level semantic prediction consistenci es at different levels of image formation. As the transfer is label-free and dat a-driven, the robustness of prediction is addressed by selectively involving a s

ubset of image regions for model regularization. Extensive experiments are conducted to verify the proposal of RPT on the transfer of models trained on GTA5 and SYNTHIA (synthetic data) to Cityscapes dataset (urban street scenes). RPT shows consistent improvements when injecting the constraints on several neural networks for semantic segmentation. More remarkably, when integrating RPT into the adversarial-based segmentation framework, we report to-date the best results: mIoU of 53.2%/51.7% when transferring from GTA5/SYNTHIA to Cityscapes, respectively.
****

Adaptive Loss-Aware Quantization for Multi-Bit Networks
Zhongnan Qu, Zimu Zhou, Yun Cheng, Lothar Thiele; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7988-7997
We investigate the compression of deep neural networks by quantizing their weights and activations into multiple binary bases, known as multi-bit networks (MBNs), which accelerate the inference and reduce the storage for the deployment on low-resource mobile and embedded platforms. We propose Adaptive Loss-aware Quantization (ALQ), a new MBN quantization pipeline that is able to achieve an average bitwidth below one-bit without notable loss in inference accuracy. Unlike previous MBN quantization solutions that train a quantizer by minimizing the error to reconstruct full precision weights, ALQ directly minimizes the quantization-induced error on the loss function involving neither gradient approximation nor full precision maintenance. ALQ also exploits strategies including adaptive bitwidth, smooth bitwidth reduction, and iterative trained quantization to allow a smaller network size without loss in accuracy. Experiment results on popular image datasets show that ALQ outperforms state-of-the-art compressed networks in terms of both storage and accuracy.
****

MaskGAN: Towards Diverse and Interactive Facial Image Manipulation
Cheng-Han Lee, Ziwei Liu, Lingyun Wu, Ping Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5549-5558
Facial image manipulation has achieved great progress in recent years. However, previous methods either operate on a predefined set of face attributes or leave users little freedom to interactively manipulate images. To overcome these drawbacks, we propose a novel framework termed MaskGAN, enabling diverse and interactive face manipulation. Our key insight is that semantic masks serve as a suitable intermediate representation for flexible face manipulation with fidelity preservation. MaskGAN has two main components: 1) Dense Mapping Network (DMN) and 2) Editing Behavior Simulated Training (EBST). Specifically, DMN learns style mapping between a free-form user modified mask and a target image, enabling diverse generation results. EBST models the user editing behavior on the source mask, making the overall framework more robust to various manipulated inputs. Specifically, it introduces dual-editing consistency as the auxiliary supervision signal. To facilitate extensive studies, we construct a large-scale high-resolution face dataset with fine-grained mask annotations named CelebAMask-HQ. MaskGAN is comprehensively evaluated on two challenging tasks: attribute transfer and style copy, demonstrating superior performance over other state-of-the-art methods. The code, models, and dataset are available at https://github.com/switchablenorms/CelebAMask-HQ.
****

ClusterFit: Improving Generalization of Visual Representations
Xueting Yan, Ishan Misra, Abhinav Gupta, Deepti Ghadiyaram, Dhruv Mahajan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6509-6518
Pre-training convolutional neural networks with weakly-supervised and self-supervised strategies is becoming increasingly popular for several computer vision tasks. However, due to the lack of strong discriminative signals, these learned representations may overfit to the pre-training objective (e.g., hashtag prediction) and not generalize well to downstream tasks. In this work, we present a simple strategy - ClusterFit to improve the robustness of the visual representations learned during pre-training. Given a dataset, we (a) cluster its features extrac

ted from a pre-trained network using k-means and (b) re-train a new network from scratch on this dataset using cluster assignments as pseudo-labels. We empirically show that clustering helps reduce the pre-training task-specific information from the extracted features thereby minimizing overfitting to the same. Our approach is extensible to different pre-training frameworks -- weak- and self-supervised, modalities -- images and videos, and pre-training tasks -- object and action classification. Through extensive transfer learning experiments on 11 different target datasets of varied vocabularies and granularities, we show that ClusterFit significantly improves the representation quality compared to the state-of-the-art large-scale (millions / billions) weakly-supervised image and video models and self-supervised image models.

****

Robust Homography Estimation via Dual Principal Component Pursuit

Tianjiao Ding, Yunchen Yang, Zhihui Zhu, Daniel P. Robinson, Rene Vidal, Laurent Kneip, Manolis C. Tsakiris; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6080-6089

We revisit robust estimation of homographies over point correspondences between two or three views, a fundamental problem in geometric vision. The analysis serves as a platform to support a rigorous investigation of Dual Principal Component Pursuit (DPCP) as a valid and powerful alternative to RANSAC for robust model fitting in multiple-view geometry. Homography fitting is cast as a robust nullspace estimation problem over either homographic or epipolar/trifocal embeddings. We prove that the nullspace of epipolar or trifocal embeddings in the homographic scenario, of dimension 3 and 6 for two and three views respectively, is defined by unique, computable homographies. Experiments show that DPCP performs on par with USAC with local optimization, while requiring an order of magnitude less computing time, and it also outperforms a recent deep learning implementation for homography estimation.

****

Face X-Ray for More General Face Forgery Detection

Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, Baining Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5001-5010

In this paper we propose a novel image representation called face X-ray for detecting forgery in face images. The face X-ray of an input face image is a greyscale image that reveals whether the input image can be decomposed into the blending of two images from different sources. It does so by showing the blending boundary for a forged image and the absence of blending for a real image. We observe that most existing face manipulation methods share a common step: blending the altered face into an existing background image. For this reason, face X-ray provides an effective way for detecting forgery generated by most existing face manipulation algorithms. Face X-ray is general in the sense that it only assumes the existence of a blending step and does not rely on any knowledge of the artifacts associated with a specific face manipulation technique. Indeed, the algorithm for computing face X-ray can be trained without fake images generated by any of the state-of-the-art face manipulation methods. Extensive experiments show that face X-ray remains effective when applied to forgery generated by unseen face manipulation techniques, while most existing face forgery detection or deepfake detection algorithms experience a significant performance drop.

****

Exploring Unlabeled Faces for Novel Attribute Discovery

Hyojin Bahng, Sunghyo Chung, Seungjoo Yoo, Jaegul Choo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5821-5830

Despite remarkable success in unpaired image-to-image translation, existing systems still require a large amount of labeled images. This is a bottleneck for their real-world applications; in practice, a model trained on labeled CelebA dataset does not work well for test images from a different distribution -- greatly limiting their application to unlabeled images of a much larger quantity. In this paper, we attempt to alleviate this necessity for labeled data in the facial im

age translation domain. We aim to explore the degree to which you can discover n
ovel attributes from unlabeled faces and perform high-quality translation. To th
is end, we use prior knowledge about the visual world as guidance to discover no
vel attributes and transfer them via a novel normalization method. Experiments s
how that our method trained on unlabeled data produces high-quality translations
, preserves identity, and be perceptually realistic, as good as, or better than,
 state-of-the-art methods trained on labeled data.
****

Spatially Attentive Output Layer for Image Classification
Ildoo Kim, Woonhyuk Baek, Sungwoong Kim; Proceedings of the IEEE/CVF Conferenc
e on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9533-9542
Most convolutional neural networks (CNNs) for image classification use a global
average pooling (GAP) followed by a fully-connected (FC) layer for output logits
. However, this spatial aggregation procedure inherently restricts the utilizati
on of location-specific information at the output layer, although this spatial i
nformation can be beneficial for classification. In this paper, we propose a nov
el spatial output layer on top of the existing convolutional feature maps to exp
licitly exploit the location-specific output information. In specific, given the
 spatial feature maps, we replace the previous GAP-FC layer with a spatially att
entive output layer (SAOL) by employing a attention mask on spatial logits. The
proposed location-specific attention selectively aggregates spatial logits withi
n a target region, which leads to not only the performance improvement but also
spatially interpretable outputs. Moreover, the proposed SAOL also permits to ful
ly exploit location-specific self-supervision as well as self-distillation to en
hance the generalization ability during training. The proposed SAOL with self-su
pervision and self-distillation can be easily plugged into existing CNNs. Experi
mental results on various classification tasks with representative architectures
 show consistent performance improvements by SAOL at almost the same computation
al cost.
****

A Shared Multi-Attention Framework for Multi-Label Zero-Shot Learning
Dat Huynh, Ehsan Elhamifar; Proceedings of the IEEE/CVF Conference on Computer
Vision and Pattern Recognition (CVPR), 2020, pp. 8776-8786
In this work, we develop a shared multi-attention model for multi-label zero-sho
t learning. We argue that designing attention mechanism for recognizing multiple
 seen and unseen labels in an image is a non-trivial task as there is no trainin
g signal to localize unseen labels and an image only contains a few present labe
ls that need attentions out of thousands of possible labels. Therefore, instead
of generating attentions for unseen labels which have unknown behaviors and coul
d focus on irrelevant regions due to the lack of any training sample, we let the
 unseen labels select among a set of shared attentions which are trained to be l
abel-agnostic and to focus on only relevant/foreground regions through our novel
 loss. Finally, we learn a compatibility function to distinguish labels based on
 the selected attention. We further propose a novel loss function that consists
of three components guiding the attention to focus on diverse and relevant image
 regions while utilizing all attention features. By extensive experiments, we sh
ow that our method improves the state of the art by 2.9% and 1.4% F1 score on th
e NUS-WIDE and the large scale Open Images datasets, respectively.
****

Optical Flow in the Dark
Yinqiang Zheng, Mingfang Zhang, Feng Lu; Proceedings of the IEEE/CVF Conferenc
e on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6749-6757
Many successful optical flow estimation methods have been proposed, but they bec
ome invalid when tested in dark scenes because low-light scenarios are not consi
dered when they are designed and current optical flow benchmark datasets lack lo
w-light samples. Even if we preprocess to enhance the dark images, which achieve
s great visual perception, it still leads to poor optical flow results or even w
orse ones, because information like motion consistency may be broken while enhan
cing. We propose an end-to-end data-driven method that avoids error accumulation
 and learns optical flow directly from low-light noisy images. Specifically, we

develop a method to synthesize large-scale low-light optical flow datasets by si
mulating the noise model on dark raw images. We also collect a new optical flow
dataset in raw format with a large range of exposure to be used as a benchmark.
The models trained on our synthetic dataset can relatively maintain optical flow
 accuracy as the image brightness descends and they outperform the existing meth
ods greatly on low-light images.
****

Painting Many Pasts: Synthesizing Time Lapse Videos of Paintings
Amy Zhao, Guha Balakrishnan, Kathleen M. Lewis, Fredo Durand, John V. Guttag
, Adrian V. Dalca; Proceedings of the IEEE/CVF Conference on Computer Vision an
d Pattern Recognition (CVPR), 2020, pp. 8435-8445
We introduce a new video synthesis task: synthesizing time lapse videos depictin
g how a given painting might have been created. Artists paint using unique combi
nations of brushes, strokes, and colors. There are often many possible ways to c
reate a given painting. Our goal is to learn to capture this rich range of possi
bilities. Creating distributions of long-term videos is a challenge for learning
-based video synthesis methods. We present a probabilistic model that, given a s
ingle image of a completed painting, recurrently synthesizes steps of the painti
ng process. We implement this model as a convolutional neural network, and intro
duce a novel training scheme to enable learning from a limited dataset of painti
ng time lapses. We demonstrate that this model can be used to sample many time s
teps, enabling long-term stochastic video synthesis. We evaluate our method on d
igital and watercolor paintings collected from video websites, and show that hum
an raters find our synthetic videos to be similar to time lapse videos produced
by real artists.
****

Learning a Neural Solver for Multiple Object Tracking
Guillem Braso, Laura Leal-Taixe; Proceedings of the IEEE/CVF Conference on Comp
uter Vision and Pattern Recognition (CVPR), 2020, pp. 6247-6257
Graphs offer a natural way to formulate Multiple Object Tracking (MOT) within th
e tracking-by-detection paradigm. However, they also introduce a major challenge
 for learning methods, as defining a model that can operate on such structured d
omain is not trivial. As a consequence, most learning-based work has been devote
d to learning better features for MOT and then using these with well-established
 optimization frameworks. In this work, we exploit the classical network flow fo
rmulation of MOT to define a fully differentiable framework based on Message Pas
sing Networks (MPNs). By operating directly on the graph domain, our method can
reason globally over an entire set of detections and predict final solutions. He
nce, we show that learning in MOT does not need to be restricted to feature extr
action, but it can also be applied to the data association step. We show a signi
ficant improvement in both MOTA and IDF1 on three publicly available benchmarks.
 Our code is available at https://bit.ly/motsolv.
****

Rethinking Data Augmentation for Image Super-resolution: A Comprehensive Analysi
s and a New Strategy
Jaejun Yoo, Namhyuk Ahn, Kyung-Ah Sohn; Proceedings of the IEEE/CVF Conference
 on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8375-8384
Data augmentation is an effective way to improve the performance of deep network
s. Unfortunately, current methods are mostly developed for high-level vision tas
ks (e.g., classification) and few are studied for low-level vision tasks (e.g.,
image restoration). In this paper, we provide a comprehensive analysis of the ex
isting augmentation methods applied to the super-resolution task. We find that t
he methods discarding or manipulating the pixels or features too much hamper the
 image restoration, where the spatial relationship is very important. Based on o
ur analyses, we propose CutBlur that cuts a low-resolution patch and pastes it t
o the corresponding high-resolution image region and vice versa. The key intuiti
on of CutBlur is to enable a model to learn not only "how" but also "where" to s
uper-resolve an image. By doing so, the model can understand "how much", instead
 of blindly learning to apply super-resolution to every given pixel. Our method
consistently and significantly improves the performance across various scenarios

, especially when the model size is big and the data is collected under real-wor
ld environments. We also show that our method improves other low-level vision ta
sks, such as denoising and compression artifact removal.
****

Evade Deep Image Retrieval by Stashing Private Images in the Hash Space
Yanru Xiao, Cong Wang, Xing Gao; Proceedings of the IEEE/CVF Conference on Com
puter Vision and Pattern Recognition (CVPR), 2020, pp. 9651-9660
With the rapid growth of visual content, deep learning to hash is gaining popula
rity in the image retrieval community recently. Although it greatly facilitates
search efficiency, privacy is also at risks when images on the web are retrieved
at a large scale and exploited as a rich mine of personal information. An adver
sary can extract private images by querying similar images from the targeted cat
egory for any usable model. Existing methods based on image processing preserve
privacy at a sacrifice of perceptual quality. In this paper, we propose a new me
chanism based on adversarial examples to "stash" private images in the deep hash
space while maintaining perceptual similarity. We first find that a simple appr
oach of hamming distance maximization is not robust against brute-force adversar
ies. Then we develop a new loss function by maximizing the hamming distance to n
ot only the original category, but also the centers from all the classes, partit
ioned into clusters of various sizes. The extensive experiment shows that the pr
oposed defense can harden the attacker's efforts by 2-7 orders of magnitude, wit
hout significant increase of computational overhead and perceptual degradation.
We also demonstrate 30-60% transferability in hash space with a black-box settin
g. The code is available at: https://github.com/sugarruy/hashstash
****

GanHand: Predicting Human Grasp Affordances in Multi-Object Scenes
Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, Grego
ry Rogez; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
Recognition (CVPR), 2020, pp. 5031-5041
The rise of deep learning has brought remarkable progress in estimating hand geo
metry from images where the hands are part of the scene. This paper focuses on a
new problem not explored so far, consisting in predicting how a human would gra
sp one or several objects, given a single RGB image of these objects. This is a
problem with enormous potential in e.g. augmented reality, robotics or prostheti
c design. In order to predict feasible grasps, we need to understand the semanti
c content of the image, its geometric structure and all potential interactions w
ith a hand physical model. To this end, we introduce a generative model that joi
ntly reasons in all these levels and 1) regresses the 3D shape and pose of the o
bjects in the scene; 2) estimates the grasp types; and 3) refines the 51-DoF of
a 3D hand model that minimize a graspability loss. To train this model we build
the YCB-Affordance dataset, that contains more than 133k images of 21 objects in
the YCB-Video dataset. We have annotated these images with more than 28M plausi
ble 3D human grasps according to a 33-class taxonomy. A thorough evaluation in s
ynthetic and real images shows that our model can robustly predict realistic gra
sps, even in cluttered scenes with multiple objects in close contact.
****

EventSR: From Asynchronous Events to Image Reconstruction, Restoration, and Supe
r-Resolution via End-to-End Adversarial Learning
Lin Wang, Tae-Kyun Kim, Kuk-Jin Yoon; Proceedings of the IEEE/CVF Conference o
n Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8315-8325
Event cameras sense intensity changes and have many advantages over conventional
cameras. To take advantage of event cameras, some methods have been proposed to
reconstruct intensity images from event streams. However, the outputs are still
in low resolution (LR), noisy, and unrealistic. The low-quality outputs stem br
oader applications of event cameras, where high spatial resolution (HR) is neede
d as well as high temporal resolution, dynamic range, and no motion blur. We con
sider the problem of reconstructing and super-resolving intensity images from pu
re events, when no ground truth (GT) HR images and down-sampling kernels are ava
ilable. To tackle the challenges, we propose a novel end-to-end pipeline that re
constructs LR images from event streams, enhances the image qualities and upsamp

les the enhanced images, called EventSR. For the absence of real GT images, our method is primarily unsupervised, deploying adversarial learning. To train EventSR, we create an open dataset including both real-world and simulated scenes. The use of both datasets boosts up the network performance, and the network architectures and various loss functions in each phase help improve the image qualities. The whole pipeline is trained in three phases. While each phase is mainly for one of the three tasks, the networks in earlier phases are fine-tuned by respective loss functions in an end-to-end manner. Experimental results show that EventSR generates high-quality SR images from events for both simulated and real-world data.

****

Quaternion Product Units for Deep Learning on 3D Rotation Groups

Xuan Zhang, Shaofei Qin, Yi Xu, Hongteng Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7304-7313

We propose a novel quaternion product unit (QPU) to represent data on 3D rotation groups. The QPU leverages quaternion algebra and the law of 3D rotation group, representing 3D rotation data as quaternions and merging them via a weighted chain of Hamilton products. We prove that the representations derived by the proposed QPU can be disentangled into "rotation-invariant" features and "rotation-equivariant" features, respectively, which supports the rationality and the efficiency of the QPU in theory. We design quaternion neural networks based on our QPUs and make our models compatible with existing deep learning models. Experiments on both synthetic and real-world data show that the proposed QPU is beneficial for the learning tasks requiring rotation robustness.

****

3D Human Mesh Regression With Dense Correspondence

Wang Zeng, Wanli Ouyang, Ping Luo, Wentao Liu, Xiaogang Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7054-7063

Estimating 3D mesh of the human body from a single 2D image is an important task with many applications such as augmented reality and Human-Robot interaction. However, prior works reconstructed 3D mesh from global image feature extracted by using convolutional neural network (CNN), where the dense correspondences between the mesh surface and the image pixels are missing, leading to suboptimal solution. This paper proposes a model-free 3D human mesh estimation framework, named DecoMR, which explicitly establishes the dense correspondence between the mesh and the local image features in the UV space (i.e. a 2D space used for texture mapping of 3D mesh). DecoMR first predicts pixel-to-surface dense correspondence map (i.e., IUV image), with which we transfer local features from the image space to the UV space. Then the transferred local image features are processed in the UV space to regress a location map, which is well aligned with transferred features. Finally we reconstruct 3D human mesh from the regressed location map with a predefined mapping function. We also observe that the existing discontinuous UV map are unfriendly to the learning of network. Therefore, we propose a novel UV map that maintains most of the neighboring relations on the original mesh surface. Experiments demonstrate that our proposed local feature alignment and continuous UV map outperforms existing 3D mesh based methods on multiple public benchmarks. Code will be made available at https: //github.com/zengwang430521/DecoMR.

****

Learning to Shadow Hand-Drawn Sketches

Qingyuan Zheng, Zhuoru Li, Adam Bargteil; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7436-7445

We present a fully automatic method to generate detailed and accurate artistic shadows from pairs of line drawing sketches and lighting directions. We also contribute a new dataset of one thousand examples of pairs of line drawings and shadows that are tagged with lighting directions. Remarkably, the generated shadows quickly communicate the underlying 3D structure of the sketched scene. Consequently, the shadows generated by our approach can be used directly or as an excellent starting point for artists. We demonstrate that the deep learning network we

propose takes a hand-drawn sketch, builds a 3D model in latent space, and renders the resulting shadows. The generated shadows respect the hand-drawn lines and underlying 3D space and contain sophisticated and accurate details, such as self-shadowing effects. Moreover, the generated shadows contain artistic effects, such as rim lighting or halos appearing from backlighting, that would be achievable with traditional 3D rendering methods.
****

Optimizing Rank-Based Metrics With Blackbox Differentiation
Michal Rolinek, Vit Musil, Anselm Paulus, Marin Vlastelica, Claudio Michaelis, Georg Martius; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7620-7630
Rank-based metrics are some of the most widely used criteria for performance evaluation of computer vision models. Despite years of effort, direct optimization for these metrics remains a challenge due to their non-differentiable and non-decomposable nature. We present an efficient, theoretically sound, and general method for differentiating rank-based metrics with mini-batch gradient descent. In addition, we address optimization instability and sparsity of the supervision signal that both arise from using rank-based metrics as optimization targets. Resulting losses based on recall and Average Precision are applied to image retrieval and object detection tasks. We obtain performance that is competitive with state-of-the-art on standard image retrieval datasets and consistently improve performance of near state-of-the-art object detectors.
****

Fast Texture Synthesis via Pseudo Optimizer
Wu Shi, Yu Qiao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5498-5507
Texture synthesis using deep neural networks can generate high quality and diversified textures. However, it usually requires a heavy optimization process. The following works accelerate the process by using feed-forward networks, but at the cost of scalability. diversity or quality. We propose a new efficient method that aims to simulate the optimization process while retains most of the properties. Our method takes a noise image and the gradients from a descriptor network as inputs, and synthesize a refined image with respect to the target image. The proposed method can synthesize images with better quality and diversity than the other fast synthesis methods do. Moreover, our method trained on a large scale dataset can generalize to synthesize unseen textures.
****

ENSEI: Efficient Secure Inference via Frequency-Domain Homomorphic Convolution for Privacy-Preserving Visual Recognition
Song Bian, Tianchen Wang, Masayuki Hiromoto, Yiyu Shi, Takashi Sato; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9403-9412
In this work, we propose ENSEI, a secure inference (SI) framework based on the frequency-domain secure convolution (FDSC) protocol for the efficient execution of image inference in the encrypted domain. Our observation is that, under the combination of homomorphic encryption and secret sharing, homomorphic convolution can be obliviously carried out in the frequency domain, significantly simplifying the related computations. We provide protocol designs and parameter derivations for number-theoretic transform (NTT) based FDSC. In the experiment, we thoroughly study the accuracy-efficiency trade-offs between time- and frequency-domain homomorphic convolution. With ENSEI, compared to the best known works, we achieve 5--11x online time reduction, up to 33x setup time reduction, and up to 10x reduction in the overall inference time. A further 33% of bandwidth reductions can be obtained on binary neural networks with only 3% of accuracy degradation on the CIFAR-10 dataset.
****

Learning Dynamic Relationships for 3D Human Motion Prediction
Qiongjie Cui, Huaijiang Sun, Fei Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6519-6527
3D human motion prediction, i.e., forecasting future sequences from given histor

ical poses, is a fundamental task for action analysis, human-computer interactio n, machine intelligence. Recently, the state-of-the-art method assumes that the whole human motion sequence involves a fully-connected graph formed by links bet ween each joint pair. Although encouraging performance has been made, due to the neglect of the inherent and meaningful characteristics of the natural connectiv ity of human joints, unexpected results may be produced. Moreover, such a compli cated topology greatly increases the training difficulty. To tackle these issues , we propose a deep generative model based on graph networks and adversarial lea rning. Specifically, the skeleton pose is represented as a novel dynamic graph, in which natural connectivities of the joint pairs are exploited explicitly, and the links of geometrically separated joints can also be learned implicitly. Not ably, in the proposed model, the natural connection strength is adaptively learn ed, whereas, in previous schemes, it was constant. Our approach is evaluated on two representations (i.e., angle-based, position-based) from various large-scale 3D skeleton benchmarks (e.g., H3.6M, CMU, 3DPW MoCap). Extensive experiments de monstrate that our approach achieves significant improvements against existing b aselines in accuracy and visualization. Code will be available at https://github .com/cuiqiongjie/LDRGCN.

****

SAM: The Sensitivity of Attribution Methods to Hyperparameters

Naman Bansal, Chirag Agarwal, Anh Nguyen; Proceedings of the IEEE/CVF Conferen ce on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8673-8683

Attribution methods can provide powerful insights into the reasons for a classif ier's decision. We argue that a key desideratum of an explanation is its robustn ess to input hyperparameter changes that are often randomly set or empirically t uned. High sensitivity to arbitrary hyperparameter choices does not only impede reproducibility but also questions the correctness of an explanation and impairs the trust by end-users. In this paper, we provide a thorough empirical study on the sensitivity of existing attribution methods. We found an alarming trend tha t many methods are highly sensitive to changes in their common hyperparameters e .g. even changing a random seed can yield a different explanation! In contrast, explanations generated for robust classifiers that are trained to be invariant t o pixel-wise perturbations are surprisingly more robust. Interestingly, such sen sitivity is not reflected in the average explanation correctness scores over the entire dataset as commonly reported in the literature.

****

Learning to Optimize on SPD Manifolds

Zhi Gao, Yuwei Wu, Yunde Jia, Mehrtash Harandi; Proceedings of the IEEE/CVF C onference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7700-7709

Many tasks in computer vision and machine learning are modeled as optimization p roblems with constraints in the form of Symmetric Positive Definite (SPD) matric es. Solving such optimization problems is challenging due to the non-linearity o f the SPD manifold, making optimization with SPD constraints heavily relying on expert knowledge and human involvement. In this paper, we propose a meta-learnin g method to automatically learn an iterative optimizer on SPD manifolds. Specifi cally, we introduce a novel recurrent model that takes into account the structur e of input gradients and identifies the updating scheme of optimization. We para meterize the optimizer by the recurrent model and utilize Riemannian operations to ensure that our method is faithful to the geometry of SPD manifolds. Compared with existing SPD optimizers, our optimizer effectively exploits the underlying data distribution and learns a better optimization trajectory in a data-driven manner. Extensive experiments on various computer vision tasks including metric nearness, clustering, and similarity learning demonstrate that our optimizer out performs existing state-of-the-art methods consistently.

****

RGBD-Dog: Predicting Canine Pose from RGBD Sensors

Sinead Kearney, Wenbin Li, Martin Parsons, Kwang In Kim, Darren Cosker; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2020, pp. 8336-8345

The automatic extraction of animal 3D pose from images without markers is of int

erest in a range of scientific fields. Most work to date predicts animal pose fr
om RGB images, based on 2D labelling of joint positions. However, due to the dif
ficult nature of obtaining training data, no ground truth dataset of 3D animal m
otion is available to quantitatively evaluate these approaches. In addition, a l
ack of 3D animal pose data also makes it difficult to train 3D pose-prediction m
ethods in a similar manner to the popular field of body-pose prediction. In our
work, we focus on the problem of 3D canine pose estimation from RGBD images, rec
ording a diverse range of dog breeds with several Microsoft Kinect v2s, simultan
eously obtaining the 3D ground truth skeleton via a motion capture system. We ge
nerate a dataset of synthetic RGBD images from this data. A stacked hourglass ne
twork is trained to predict 3D joint locations, which is then constrained using
prior models of shape and pose. We evaluate our model on both synthetic and real
 RGBD images and compare our results to previously published work fitting canine
 models to images. Finally, despite our training set consisting only of dog data
, visual inspection implies that our network can produce good predictions for im
ages of other quadrupeds - e.g. horses or cats - when their pose is similar to t
hat contained in our training set.
****

CookGAN: Causality Based Text-to-Image Synthesis
Bin Zhu, Chong-Wah Ngo; Proceedings of the IEEE/CVF Conference on Computer Visi
on and Pattern Recognition (CVPR), 2020, pp. 5519-5527
This paper addresses the problem of text-to-image synthesis from a new perspecti
ve, i.e., the cause-and-effect chain in image generation. Causality is a common
phenomenon in cooking. The dish appearance changes depending on the cooking acti
ons and ingredients. The challenge of synthesis is that a generated image should
 depict the visual result of action-on-object. This paper presents a new network
 architecture, CookGAN, that mimics visual effect in causality chain, preserves
fine-grained details and progressively upsamples image. Particularly, a cooking
simulator sub-network is proposed to incrementally make changes to food images b
ased on the interaction between ingredients and cooking methods over a series of
 steps. Experiments on Recipe1M verify that CookGAN manages to generate food ima
ges with reasonably impressive inception score. Furthermore, the images are sema
ntically interpretable and manipulable.
****

Image Based Virtual Try-On Network From Unpaired Data
Assaf Neuberger, Eran Borenstein, Bar Hilleli, Eduard Oks, Sharon Alpert; Pr
oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
(CVPR), 2020, pp. 5184-5193
This paper presents a new image-based virtual try-on approach (Outfit-VITON) tha
t helps visualize how a composition of clothing items selected from various refe
rence images form a cohesive outfit on a person in a query image. Our algorithm
has two distinctive properties. First, it is inexpensive, as it simply requires
a large set of single (non-corresponding) images (both real and catalog) of peop
le wearing various garments without explicit 3D information. The training phase
requires only single images, eliminating the need for manually creating image pa
irs, where one image shows a person wearing a particular garment and the other s
hows the same catalog garment alone. Secondly, it can synthesize images of multi
ple garments composed into a single, coherent outfit; and it enables control of
the type of garments rendered in the final outfit. Once trained, our approach ca
n then synthesize a cohesive outfit from multiple images of clothed human models
, while fitting the outfit to the body shape and pose of the query person. An on
line optimization step takes care of fine details such as intricate textures and
 logos. Quantitative and qualitative evaluations on an image dataset containing
large shape and style variations demonstrate superior accuracy compared to exist
ing state-of-the-art methods, especially when dealing with highly detailed garme
nts.
****

EventCap: Monocular 3D Capture of High-Speed Human Motions Using an Event Camera
Lan Xu, Weipeng Xu, Vladislav Golyanik, Marc Habermann, Lu Fang, Christian
Theobalt; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern

Recognition (CVPR), 2020, pp. 4968-4978
The high frame rate is a critical requirement for capturing fast human motions. In this setting, existing markerless image-based methods are constrained by the lighting requirement, the high data bandwidth and the consequent high computation overhead. In this paper, we propose EventCap -- the first approach for 3D capturing of high-speed human motions using a single event camera. Our method combines model-based optimization and CNN-based human pose detection to capture high frequency motion details and to reduce the drifting in the tracking. As a result, we can capture fast motions at millisecond resolution with significantly higher data efficiency than using high frame rate videos. Experiments on our new event-based fast human motion dataset demonstrate the effectiveness and accuracy of our method, as well as its robustness to challenging lighting conditions.
****

Dreaming to Distill: Data-Free Knowledge Transfer via DeepInversion
Hongxu Yin, Pavlo Molchanov, Jose M. Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K. Jha, Jan Kautz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8715-8724
We introduce DeepInversion, a new method for synthesizing images from the image distribution used to train a deep neural network. We "invert" a trained network (teacher) to synthesize class-conditional input images starting from random noise, without using any additional information about the training dataset. Keeping the teacher fixed, our method optimizes the input while regularizing the distribution of intermediate feature maps using information stored in the batch normalization layers of the teacher. Further, we improve the diversity of synthesized images using Adaptive DeepInversion, which maximizes the Jensen-Shannon divergence between the teacher and student network logits. The resulting synthesized images from networks trained on the CIFAR-10 and ImageNet datasets demonstrate high fidelity and degree of realism, and help enable a new breed of data-free applications - ones that do not require any real images or labeled data. We demonstrate the applicability of our proposed method to three tasks of immense practical importance - (i) data-free network pruning, (ii) data-free knowledge transfer, and (iii) data-free continual learning.
****

Spherical Space Domain Adaptation With Robust Pseudo-Label Loss
Xiang Gu, Jian Sun, Zongben Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9101-9110
Adversarial domain adaptation (DA) has been an effective approach for learning domain-invariant features by adversarial training. In this paper, we propose a novel adversarial DA approach completely defined in spherical feature space, in which we define spherical classifier for label prediction and spherical domain discriminator for discriminating domain labels. To utilize pseudo-label robustly, we develop a robust pseudo-label loss in the spherical feature space, which weights the importance of estimated labels of target data by posterior probability of correct labeling, modeled by Gaussian-uniform mixture model in spherical feature space. Extensive experiments show that our method achieves state-of-the-art results, and also confirm effectiveness of spherical classifier, spherical discriminator and spherical robust pseudo-label loss.
****

Approximating shapes in images with low-complexity polygons
Muxingzi Li, Florent Lafarge, Renaud Marlet; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8633-8641
We present an algorithm for extracting and vectorizing objects in images with polygons. Departing from a polygonal partition that oversegments an image into convex cells, the algorithm refines the geometry of the partition while labeling its cells by a semantic class. The result is a set of polygons, each capturing an object in the image. The quality of a configuration is measured by an energy that accounts for both the fidelity to input data and the complexity of the output polygons. To efficiently explore the configuration space, we perform splitting and merging operations in tandem on the cells of the polygonal partition. The exploration mechanism is controlled by a priority queue that sorts the operations m

ost likely to decrease the energy. We show the potential of our algorithm on dif ferent types of scenes, from organic shapes to man-made objects through floor ma ps, and demonstrate its efficiency compared to existing vectorization methods.
****

Vec2Face: Unveil Human Faces From Their Blackbox Features in Face Recognition
Chi Nhan Duong, Thanh-Dat Truong, Khoa Luu, Kha Gia Quach, Hung Bui, Kaushi k Roy; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec ognition (CVPR), 2020, pp. 6132-6141
Unveiling face images of a subject given his/her high-level representations extr acted from a blackbox Face Recognition engine is extremely challenging. It is be cause the limitations of accessible information from that engine including its s tructure and uninterpretable extracted features. This paper presents a novel gen erative structure with Bijective Metric Learning, namely Bijective Generative Ad versarial Networks in a Distillation framework (DiBiGAN), for synthesizing faces of an identity given that person's features. In order to effectively address th is problem, this work firstly introduces a bijective metric so that the distance measurement and metric learning process can be directly adopted in image domain for an image reconstruction task. Secondly, a distillation process is introduce d to maximize the information exploited from the blackbox face recognition engin e. Then a Feature-Conditional Generator Structure with Exponential Weighting Str ategy is presented for a more robust generator that can synthesize realistic fac es with ID preservation. Results on several benchmarking datasets including Cele bA, LFW, AgeDB, CFP-FP against matching engines have demonstrated the effectiven ess of DiBiGAN on both image realism and ID preservation properties.
****

SiamCAR: Siamese Fully Convolutional Classification and Regression for Visual Tr acking
Dongyan Guo, Jun Wang, Ying Cui, Zhenhua Wang, Shengyong Chen; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202 0, pp. 6269-6277
By decomposing the visual tracking task into two subproblems as classification f or pixel category and regression for object bounding box at this pixel, we propo se a novel fully convolutional Siamese network to solve visual tracking end-to-e nd in a per-pixel manner. The proposed framework SiamCAR consists of two simple subnetworks: one Siamese subnetwork for feature extraction and one classificatio n-regression subnetwork for bounding box prediction. Different from state-of-the -art trackers like Siamese-RPN, SiamRPN++ and SPM, which are based on region pro posal, the proposed framework is both proposal and anchor free. Consequently, we are able to avoid the tricky hyper-parameter tuning of anchors and reduce human intervention. The proposed framework is simple, neat and effective. Extensive e xperiments and comparisons with state-of-the-art trackers are conducted on chall enging benchmarks including GOT-10K, LaSOT, UAV123 and OTB-50. Without bells and whistles, our SiamCAR achieves the leading performance with a considerable real -time speed. The code is available at https://github.com/ohhhyeahhh/SiamCAR.
****

Deep Image Spatial Transformation for Person Image Generation
Yurui Ren, Xiaoming Yu, Junming Chen, Thomas H. Li, Ge Li; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, p p. 7690-7699
Pose-guided person image generation is to transform a source person image to a t arget pose. This task requires spatial manipulations of source data. However, Co nvolutional Neural Networks are limited by the lack of ability to spatially tran sform the inputs. In this paper, we propose a differentiable global-flow local-a ttention framework to reassemble the inputs at the feature level. Specifically, our model first calculates the global correlations between sources and targets t o predict flow fields. Then, the flowed local patch pairs are extracted from the feature maps to calculate the local attention coefficients. Finally, we warp th e source features using a content-aware sampling method with the obtained local attention coefficients. The results of both subjective and objective experiments demonstrate the superiority of our model. Besides, additional results in video

animation and view synthesis show that our model is applicable to other tasks re quiring spatial transformation. Our source code is available at https://github.c om/RenYurui/Global-Flow-Local-Attention.
****

Fashion Editing With Adversarial Parsing Learning
Haoye Dong, Xiaodan Liang, Yixuan Zhang, Xujie Zhang, Xiaohui Shen, Zhenyu Xie, Bowen Wu, Jian Yin; Proceedings of the IEEE/CVF Conference on Computer Vi sion and Pattern Recognition (CVPR), 2020, pp. 8120-8128
Interactive fashion image manipulation, which enables users to edit images with sketches and color strokes, is an interesting research problem with great applic ation value. Existing works often treat it as a general inpainting task and do n ot fully leverage the semantic structural information in fashion images. Moreove r, they directly utilize conventional convolution and normalization layers to re store the incomplete image, which tends to wash away the sketch and color inform ation. In this paper, we propose a novel Fashion Editing Generative Adversarial Network (FE-GAN), which is capable of manipulating fashion images by free-form s ketches and sparse color strokes. FE-GAN consists of two modules: 1) a free-form parsing network that learns to control the human parsing generation by manipula ting sketch and color; 2) a parsing-aware inpainting network that renders detail ed textures with semantic guidance from the human parsing map. A new attention n ormalization layer is further applied at multiple scales in the decoder of the i npainting network to enhance the quality of the synthesized image. Extensive exp eriments on high-resolution fashion image datasets demonstrate that the proposed FE-GAN significantly outperforms the state-of-the-art methods on fashion image manipulation.
****

Multiview-Consistent Semi-Supervised Learning for 3D Human Pose Estimation
Rahul Mitra, Nitesh B. Gundavarapu, Abhishek Sharma, Arjun Jain; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20 20, pp. 6907-6916
The best performing methods for 3D human pose estimation from monocular images r equire large amounts of in-the-wild 2D and controlled 3D pose annotated datasets which are costly and require sophisticated systems to acquire. To reduce this a nnotation dependency, we propose Multiview-Consistent Semi Supervised Learning (MCSS) framework that utilizes similarity in pose information from unannotated, u ncalibrated but synchronized multi-view videos of human motions as additional we ak supervision signal to guide 3D human pose regression. Our framework applies h ard-negative mining based on temporal relations in multi-view videos to arrive a t a multi-view consistent pose embedding and when jointly trained with limited 3 D pose annotations, our approach improves the baseline by 25% and state-of-the-a rt by 8.7%, whilst using substantially smaller networks. Lastly, but importantly , we demonstrate the advantages of the learned embedding and establish view-inva riant pose retrieval benchmarks on two popular, publicly available multi-view hu man pose datasets, Human 3.6M and MPI-INF-3DHP, to facilitate future research.
****

Attack to Explain Deep Representation
Mohammad A. A. K. Jalwana, Naveed Akhtar, Mohammed Bennamoun, Ajmal Mian; Pro ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9543-9552
Deep visual models are susceptible to extremely low magnitude perturbations to i nput images. Though carefully crafted, the perturbation patterns generally appea r noisy, yet they are able to perform controlled manipulation of model predictio ns. This observation is used to argue that deep representation is misaligned wit h human perception. This paper counter-argues and proposes the first attack on d eep learning that aims at explaining the learned representation instead of fooli ng it. By extending the input domain of the manipulative signal and employing a model faithful channelling, we iteratively accumulate adversarial perturbations for a deep model. The accumulated signal gradually manifests itself as a collect ion of visually salient features of the target label (in model fooling), casting adversarial perturbations as primitive features of the target label. Our attack

provides the first demonstration of systematically computing perturbations for adversarially non-robust classifiers that comprise salient visual features of objects. We leverage the model explaining character of our algorithm to perform image generation, inpainting and interactive image manipulation by attacking adversarially robust classifiers. The visually appealing results across these applications demonstrate the utility of our attack (and perturbations in general) beyond model fooling.

****

FALCON: A Fourier Transform Based Approach for Fast and Secure Convolutional Neural Network Predictions

Shaohua Li, Kaiping Xue, Bin Zhu, Chenkai Ding, Xindi Gao, David Wei, Tao Wan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8705-8714

Deep learning as a service has been widely deployed to utilize deep neural network models to provide prediction services. However, this raises privacy concerns since clients need to send sensitive information to servers. In this paper, we focus on the scenario where clients want to classify private images with a convolutional neural network model hosted in the server, while both parties keep their data private. We present FALCON, a fast and secure approach for CNN predictions based on fast Fourier Transform. Our solution enables linear layers of a CNN model to be evaluated simply and efficiently with fully homomorphic encryption. We also introduce the first efficient and privacy-preserving protocol for softmax function, which is an indispensable component in CNNs and has not yet been evaluated in previous work due to its high complexity.

****

The Knowledge Within: Methods for Data-Free Model Compression

Matan Haroush, Itay Hubara, Elad Hoffer, Daniel Soudry; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8494-8502

Background: Recently, an extensive amount of research has been focused on compressing and accelerating Deep Neural Networks (DNN). So far, high compression rate algorithms require part of the training dataset for a low precision calibration, or a fine-tuning process. However, this requirement is unacceptable when the data is unavailable or contains sensitive information, as in medical and biometric use-cases. Contributions: We present three methods for generating synthetic samples from trained models. Then, we demonstrate how these samples can be used to calibrate and fine-tune quantized models without using any real data in the process. Our best performing method has a negligible accuracy degradation compared to the original training set. This method, which leverages intrinsic batch normalization layers' statistics of the trained model, can be used to evaluate data similarity. Our approach opens a path towards genuine data-free model compression, alleviating the need for training data during model deployment.

****

PropagationNet: Propagate Points to Curve to Learn Structure Information

Xiehe Huang, Weihong Deng, Haifeng Shen, Xiubao Zhang, Jieping Ye; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7265-7274

Deep learning technique has dramatically boosted the performance of face alignment algorithms. However, due to large variability and lack of samples, the alignment problem in unconstrained situations, e.g. large head poses, exaggerated expression, and uneven illumination, is still largely unsolved. In this paper, we explore the instincts and reasons behind our two proposals, i.e. Propagation Module and Focal Wing Loss, to tackle the problem. Concretely, we present a novel structure-infused face alignment algorithm based on heatmap regression via propagating landmark heatmaps to boundary heatmaps, which provide structure information for further attention map generation. Moreover, we propose a Focal Wing Loss for mining and emphasizing the difficult samples under in-the-wild condition. In addition, we adopt methods like CoordConv and Anti-aliased CNN from other fields that address the shift variance problem of CNN for face alignment. When implementing extensive experiments on different benchmarks, i.e. WFLW, 300W, and COFW, ou

r method outperforms the state-of-the-arts by a significant margin. Our proposed approach achieves 4.05% mean error on WFLW, 2.93% mean error on 300W full-set, and 3.71% mean error on COFW.
****

S3VAE: Self-Supervised Sequential VAE for Representation Disentanglement and Data Generation
Yizhe Zhu, Martin Renqiang Min, Asim Kadav, Hans Peter Graf; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6538-6547
We propose a sequential variational autoencoder to learn disentangled representations of sequential data (e.g., videos and audios) under self-supervision. Specifically, we exploit the benefits of some readily accessible supervision signals from input data itself or some off-the-shelf functional models and accordingly design auxiliary tasks for our model to utilize these signals. With the supervision of the signals, our model can easily disentangle the representation of an input sequence into static factors and dynamic factors (i.e., time-invariant and time-varying parts). Comprehensive experiments across videos and audios verify the effectiveness of our model on representation disentanglement and generation of sequential data, and demonstrate that, our model with self-supervision performs comparable to, if not better than, the fully-supervised model with ground truth labels, and outperforms state-of-the-art unsupervised models by a large margin.
****

Same Features, Different Day: Weakly Supervised Feature Learning for Seasonal Invariance
Jaime Spencer, Richard Bowden, Simon Hadfield; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6459-6468
"Like night and day" is a commonly used expression to imply that two things are completely different. Unfortunately, this tends to be the case for current visual feature representations of the same scene across varying seasons or times of day. The aim of this paper is to provide a dense feature representation that can be used to perform localization, sparse matching or image retrieval, regardless of the current seasonal or temporal appearance. Recently, there have been several proposed methodologies for deep learning dense feature representations. These methods make use of ground truth pixel-wise correspondences between pairs of images and focus on the spatial properties of the features. As such, they don't address temporal or seasonal variation. Furthermore, obtaining the required pixel-wise correspondence data to train in cross-seasonal environments is highly complex in most scenarios. We propose Deja-Vu, a weakly supervised approach to learning season invariant features that does not require pixel-wise ground truth data. The proposed system only requires coarse labels indicating if two images correspond to the same location or not. From these labels, the network is trained to produce "similar" dense feature maps for corresponding locations despite environmental changes. Code will be made available at: https://github.com/jspenmar/DejaVu_Features
****

Implicit Functions in Feature Space for 3D Shape Reconstruction and Completion
Julian Chibane, Thiemo Alldieck, Gerard Pons-Moll; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6970-6981
While many works focus on 3D reconstruction from images, in this paper, we focus on 3D shape reconstruction and completion from a variety of 3D inputs, which are deficient in some respect: low and high resolution voxels, sparse and dense point clouds, complete or incomplete. Processing of such 3D inputs is an increasingly important problem as they are the output of 3D scanners, which are becoming more accessible, and are the intermediate output of 3D computer vision algorithms. Recently, learned implicit functions have shown great promise as they produce continuous reconstructions. However, we identified two limitations in reconstruction from 3D inputs: 1) details present in the input data are not retained, and 2) poor reconstruction of articulated humans. To solve this, we propose Implicit Feature Networks (IF-Nets), which deliver continuous outputs, can handle multi

ple topologies, and complete shapes for missing or sparse input data retaining t
he nice properties of recent learned implicit functions, but critically they can
 also retain detail when it is present in the input data, and can reconstruct ar
ticulated humans. Our work differs from prior work in two crucial aspects. First
, instead of using a single vector to encode a 3D shape, we extract a learnable
3-dimensional multi-scale tensor of deep features, which is aligned with the ori
ginal Euclidean space embedding the shape. Second, instead of classifying x-y-z
point coordinates directly, we classify deep features extracted from the tensor
at a continuous query point. We show that this forces our model to make decision
s based on global and local shape structure, as opposed to point coordinates, wh
ich are arbitrary under Euclidean transformations. Experiments demonstrate that
IF-Nets outperform prior work in 3D object reconstruction in ShapeNet, and obtai
n significantly more accurate 3D human reconstructions. Code and project website
 is available at https://virtualhumans.mpi-inf.mpg.de/ifnets/.
****

AdaCoSeg: Adaptive Shape Co-Segmentation With Group Consistency Loss
Chenyang Zhu, Kai Xu, Siddhartha Chaudhuri, Li Yi, Leonidas J. Guibas, Hao
Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec
ognition (CVPR), 2020, pp. 8543-8552
We introduce AdaCoSeg, a deep neural network architecture for adaptive co-segmen
tation of a set of 3D shapes represented as point clouds. Differently from the f
amiliar single-instance segmentation problem, co-segmentation is intrinsically c
ontextual: how a shape is segmented can vary depending on the set it is in. Henc
e, our network features an adaptive learning module to produce a consistent shap
e segmentation which adapts to a set. Specifically, given an input set of unsegm
ented shapes, we first employ an offline pre-trained part prior network to propo
se per-shape parts. Then the co-segmentation network iteratively and jointly opt
imizes the part labelings across the set subjected to a novel group consistency
loss defined by matrix ranks. While the part prior network can be trained with n
oisy and inconsistently segmented shapes, the final output of AdaSeg is a consis
tent part labeling for the input set, with each shape segmented into up to (a us
er-specified) K parts. Overall, our method is weakly supervised, producing segme
ntations tailored to the test set, without consistent ground-truth segmentations
. We show qualitative and quantitative results from AdaSeg and evaluate it via a
blation studies and comparisons to state-of-the-art co-segmentation methods.
****

Learning Combinatorial Solver for Graph Matching
Tao Wang, He Liu, Yidong Li, Yi Jin, Xiaohui Hou, Haibin Ling; Proceedings
of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20
20, pp. 7568-7577
Learning-based approaches to graph matching have been developed and explored for
 more than a decade, have grown rapidly in scope and popularity in recent years.
 However, previous learning-based algorithms, with or without deep learning stra
tegy, mainly focus on the learning of node and/or edge affinities generation, an
d pay less attention on the learning of the combinatorial solver. In this paper
we propose a fully trainable framework for graph matching, in which learning of
affinities and solving for combinatorial optimization are not explicitly separat
ed as in many previous arts. We firstly convert the problem of building node cor
respondences between two input graphs to the problem of selecting reliable nodes
 from a constructed assignment graph. Subsequently, the graph network block modu
le is adopted to perform computation on the graph to form structured representat
ions for each node. It finally predicts a label for each node that is used for n
ode classification, and the training is performed under the supervision of both
permutation differences and the one-to-one matching constraints. The proposed me
thod is evaluated on four public benchmarks in comparison with several state-of-
the-art algorithms, and the experimental results illustrate its excellent perfor
mance.
****

Nonparametric Object and Parts Modeling With Lie Group Dynamics
David S. Hayden, Jason Pacheco, John W. Fisher III; Proceedings of the IEEE/CV

F Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7426-7435
Articulated motion analysis often utilizes strong prior knowledge such as a known or trained parts model for humans. Yet, the world contains a variety of articulating objects--mammals, insects, mechanized structures--where the number and configuration of parts for a particular object is unknown in advance. Here, we relax such strong assumptions via an unsupervised, Bayesian nonparametric parts model that infers an unknown number of parts with motions coupled by a body dynamic and parameterized by SE(D), the Lie group of rigid transformations. We derive an inference procedure that utilizes short observation sequences (image, depth, point cloud or mesh) of an object in motion without need for markers or learned body models. Efficient Gibbs decompositions for inference over distributions on SE(D) demonstrate robust part decompositions of moving objects under both 3D and 2D observation models. The inferred representation permits novel analysis, such as object segmentation by relative part motion, and transfers to new observations of the same object type.
****

A Neural Rendering Framework for Free-Viewpoint Relighting
Zhang Chen, Anpei Chen, Guli Zhang, Chengyuan Wang, Yu Ji, Kiriakos N. Kutulakos, Jingyi Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5599-5610
We present a novel Relightable Neural Renderer (RNR) for simultaneous view synthesis and relighting using multi-view image inputs. Existing neural rendering (NR) does not explicitly model the physical rendering process and hence has limited capabilities on relighting. RNR instead models image formation in terms of environment lighting, object intrinsic attributes, and light transport function (LTF), each corresponding to a learnable component. In particular, the incorporation of a physically based rendering process not only enables relighting but also improves the quality of view synthesis. Comprehensive experiments on synthetic and real data show that RNR provides a practical and effective solution for conducting free-viewpoint relighting.
****

Attribution in Scale and Space
Shawn Xu, Subhashini Venugopalan, Mukund Sundararajan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9680-9689
We study the attribution problem for deep networks applied to perception tasks. For vision tasks, attribution techniques attribute the prediction of a network to the pixels of the input image. We propose a new technique called Blur Integrated Gradients (Blur IG). This technique has several advantages over other methods. First, it can tell at what scale a network recognizes an object. It produces scores in the scale/frequency dimension, that we find captures interesting phenomena. Second, it satisfies the scale-space axioms, which imply that it employs perturbations that are free of artifact. We therefore produce explanations that are cleaner and consistent with the operation of deep networks. Third, it eliminates the need for baseline parameter for Integrated Gradients for perception tasks. This is desirable because the choice of baseline has a significant effect on the explanations. We compare the proposed technique against previous techniques and demonstrate application on three tasks: ImageNet object recognition, Diabetic Retinopathy prediction, and AudioSet audio event identification. Code and examples are at https://github.com/PAIR-code/saliency.
****

Probabilistic Regression for Visual Tracking
Martin Danelljan, Luc Van Gool, Radu Timofte; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7183-7192
Visual tracking is fundamentally the problem of regressing the state of the target in each video frame. While significant progress has been achieved, trackers are still prone to failures and inaccuracies. It is therefore crucial to represent the uncertainty in the target estimation. Although current prominent paradigms rely on estimating a state-dependent confidence score, this value lacks a clear

probabilistic interpretation, complicating its use. In this work, we therefore propose a probabilistic regression formulation and apply it to tracking. Our network predicts the conditional probability density of the target state given an input image. Crucially, our formulation is capable of modeling label noise stemming from inaccurate annotations and ambiguities in the task. The regression network is trained by minimizing the Kullback-Leibler divergence. When applied for tracking, our formulation not only allows a probabilistic representation of the output, but also substantially improves the performance. Our tracker sets a new state-of-the-art on six datasets, achieving 59.8% AUC on LaSOT and 75.8% Success on TrackingNet. The code and models are available at https://github.com/visionml/pytracking.

****

3DRegNet: A Deep Neural Network for 3D Point Registration

G. Dias Pais, Srikumar Ramalingam, Venu Madhav Govindu, Jacinto C. Nascimento, Rama Chellappa, Pedro Miraldo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7193-7203

We present 3DRegNet, a novel deep learning architecture for the registration of 3D scans. Given a set of 3D point correspondences, we build a deep neural network to address the following two challenges: (i) classification of the point correspondences into inliers/outliers, and (ii) regression of the motion parameters that align the scans into a common reference frame. With regard to regression, we present two alternative approaches: (i) a Deep Neural Network (DNN) registration and (ii) a Procrustes approach using SVD to estimate the transformation. Our correspondence-based approach achieves a higher speedup compared to competing baselines. We further propose the use of a refinement network, which consists of a smaller 3DRegNet as a refinement to improve the accuracy of the registration. Extensive experiments on two challenging datasets demonstrate that we outperform other methods and achieve state-of-the-art results. The code is available.

****

SEAN: Image Synthesis With Semantic Region-Adaptive Normalization

Peihao Zhu, Rameen Abdal, Yipeng Qin, Peter Wonka; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5104-5113

We propose semantic region-adaptive normalization (SEAN), a simple but effective building block for Generative Adversarial Networks conditioned on segmentation masks that describe the semantic regions in the desired output image. Using SEAN normalization, we can build a network architecture that can control the style of each semantic region individually, e.g., we can specify one style reference image per region. SEAN is better suited to encode, transfer, and synthesize style than the best previous method in terms of reconstruction quality, variability, and visual quality. We evaluate SEAN on multiple datasets and report better quantitative metrics (e.g. FID, PSNR) than the current state of the art. SEAN also pushes the frontier of interactive image editing. We can interactively edit images by changing segmentation masks or the style for any given region. We can also interpolate styles from two reference images per region.

****

Robust Reference-Based Super-Resolution With Similarity-Aware Deformable Convolution

Gyumin Shim, Jinsun Park, In So Kweon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8425-8434

In this paper, we propose a novel and efficient reference feature extraction module referred to as the Similarity Search and Extraction Network (SSEN) for reference-based super-resolution (RefSR) tasks. The proposed module extracts aligned relevant features from a reference image to increase the performance over single image super-resolution (SISR) methods. In contrast to conventional algorithms which utilize brute-force searches or optical flow estimations, the proposed algorithm is end-to-end trainable without any additional supervision or heavy computation, predicting the best match with a single network forward operation. Moreover, the proposed module is aware of not only the best matching position but also the relevancy of the best match. This makes our algorithm substantially robust

when irrelevant reference images are given, overcoming the major cause of the performance degradation when using existing RefSR methods. Furthermore, our module can be utilized for self-similarity SR if no reference image is available. Experimental results demonstrate the superior performance of the proposed algorithm compared to previous works both quantitatively and qualitatively.
****

Search to Distill: Pearls Are Everywhere but Not the Eyes
Yu Liu, Xuhui Jia, Mingxing Tan, Raviteja Vemulapalli, Yukun Zhu, Bradley Green, Xiaogang Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7539-7548
Standard Knowledge Distillation (KD) approaches distill the knowledge of a cumbersome teacher model into the parameters of a student model with a pre-defined architecture. However, the knowledge of a neural network, which is represented by the network's output distribution conditioned on its input, depends not only on its parameters but also on its architecture. Hence, a more generalized approach for KD is to distill the teacher's knowledge into both the parameters and architecture of the student. To achieve this, we present a new Architecture-aware Knowledge Distillation (AKD) approach that finds student models (pearls for the teacher) that are best for distilling the given teacher model. In particular, we leverage Neural Architecture Search (NAS), equipped with our KD-guided reward, to search for the best student architectures for a given teacher. Experimental results show our proposed AKD consistently outperforms the conventional NAS plus KD approach, and achieves state-of-the-art results on the ImageNet classification task under various latency settings. Furthermore, the best AKD student architecture for the ImageNet classification task also transfers well to other tasks such as million level face recognition and ensemble learning.
****

Boosting Semantic Human Matting With Coarse Annotations
Jinlin Liu, Yuan Yao, Wendi Hou, Miaomiao Cui, Xuansong Xie, Changshui Zhang, Xian-Sheng Hua; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8563-8572
Semantic human matting aims to estimate the per-pixel opacity of the foreground human regions. It is quite challenging that usually requires user interactive trimaps and plenty of high quality annotated data. Annotating such kind of data is labor intensive and requires great skills beyond normal users, especially considering the very detailed hair part of humans. In contrast, coarse annotated human dataset is much easier to acquire and collect from the public dataset. In this paper, we propose to leverage coarse annotated data coupled with fine annotated data to boost end-to-end semantic human matting without trimaps as extra input. Specifically, We train a mask prediction network to estimate the coarse semantic mask using the hybrid data, and then propose a quality unification network to unify the quality of the previous coarse mask outputs. A matting refinement network takes the unified mask and the input image to predict the final alpha matte. The collected coarse annotated dataset enriches our dataset significantly, allows generating high quality alpha matte for real images. Experimental results show that the proposed method performs comparably against state-of-the-art methods. Moreover, the proposed method can be used for refining coarse annotated public dataset, as well as semantic segmentation methods, which reduces the cost of annotating high quality human data to a great extent.
****

Few-Shot Learning via Embedding Adaptation With Set-to-Set Functions
Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, Fei Sha; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8808-8817
Learning with limited data is a key challenge for visual recognition. Many few-shot learning methods address this challenge by learning an instance embedding function from seen classes and apply the function to instances from unseen classes with limited labels. This style of transfer learning is task-agnostic: the embedding function is not learned optimally discriminative with respect to the unseen classes, where discerning among them leads to the target task. In this paper, we propose a novel approach to adapt the instance embeddings to the target class

ification task with a set-to-set function, yielding embeddings that are task-specific and are discriminative. We empirically investigated various instantiations of such set-to-set functions and observed the Transformer is most effective --- as it naturally satisfies key properties of our desired model. We denote this model as FEAT (few-shot embedding adaptation w/ Transformer) and validate it on both the standard few-shot classification benchmark and four extended few-shot learning settings with essential use cases, i.e., cross-domain, transductive, generalized few-shot learning, and low-shot learning. It archived consistent improvements over baseline models as well as previous methods, and established the new state-of-the-art results on two benchmarks.
****

FM2u-Net: Face Morphological Multi-Branch Network for Makeup-Invariant Face Verification
Wenxuan Wang, Yanwei Fu, Xuelin Qian, Yu-Gang Jiang, Qi Tian, Xiangyang Xue ; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5730-5740
It is challenging in learning a makeup-invariant face verification model, due to (1) insufficient makeup/non-makeup face training pairs, (2) the lack of diverse makeup faces, and (3) the significant appearance changes caused by cosmetics. To address these challenges, we propose a unified Face Morphological Multi-branch Network (FMMu-Net) for makeup-invariant face verification, which can simultaneously synthesize many diverse makeup faces through face morphology network (FM-Net) and effectively learn cosmetics-robust face representations using attention-based multi-branch learning network (AttM-Net). For challenges (1) and (2), FM-Net (two stacked auto-encoders) can synthesize realistic makeup face images by transferring specific regions of cosmetics via cycle consistent loss. For challenge (3), AttM-Net, consisting of one global and three local (task-driven on two eyes and mouth) branches, can effectively capture the complementary holistic and detailed information. Unlike DeepID2 which uses simple concatenation fusion, we introduce a heuristic method AttM-FM, attached to AttM-Net, to adaptively weight the features of different branches guided by the holistic information. We conduct extensive experiments on makeup face verification benchmarks (M-501, M-203, and FAM) and general face recognition datasets (LFW and IJB-A). Our framework FMMu-Net achieves state-of-the-art performances.
****

Deep Semantic Clustering by Partition Confidence Maximisation
Jiabo Huang, Shaogang Gong, Xiatian Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8849-8858
By simultaneously learning visual features and data grouping, deep clustering has shown impressive ability to deal with unsupervised learning for structure analysis of high-dimensional visual data. Existing deep clustering methods typically rely on local learning constraints based on inter-sample relations and/or self-estimated pseudo labels. This is susceptible to the inevitable errors distributed in the neighbourhoods and suffers from error-propagation during training. In this work, we propose to solve this problem by learning the most confident clustering solution from all the possible separations, based on the observation that assigning samples from the same semantic categories into different clusters will reduce both the intra-cluster compactness and inter-cluster diversity, i.e. lower partition confidence. Specifically, we introduce a novel deep clustering method named PartItion Confidence mAximisation (PICA). It is established on the idea of learning the most semantically plausible data separation, in which all clusters can be mapped to the ground-truth classes one-to-one, by maximising the "global" partition confidence of clustering solution. This is realised by introducing a differentiable partition uncertainty index and its stochastic approximation as well as a principled objective loss function that minimises such index, all of which together enables a direct adoption of the conventional deep networks and mini-batch based model training. Extensive experiments on six widely-adopted clustering benchmarks demonstrate our model's performance superiority over a wide range of the state-of-the-art approaches. The code is available online.
****

A Transductive Approach for Video Object Segmentation
Yizhuo Zhang, Zhirong Wu, Houwen Peng, Stephen Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6949-6958
Semi-supervised video object segmentation aims to separate a target object from a video sequence, given the mask in the first frame. Most of current prevailing methods utilize information from additional modules trained in other domains like optical flow and instance segmentation, and as a result they do not compete with other methods on common ground. To address this issue, we propose a simple yet strong transductive method, in which additional modules, datasets, and dedicated architectural designs are not needed. Our method takes a label propagation approach where pixel labels are passed forward based on feature similarity in an embedding space. Different from other propagation methods, ours diffuses temporal information in a holistic manner which take accounts of long-term object appearance. In addition, our method requires few additional computational overhead, and runs at a fast 37 fps speed. Our single model with a vanilla ResNet50 backbone achieves an overall score of 72.3% on the DAVIS 2017 validation set and 63.1% on the test set. This simple yet high performing and efficient method can serve as a solid baseline that facilitates future research. Code and models are available at https://github.com/ microsoft/transductive-vos.pytorch.
****

Uncertainty-Aware Mesh Decoder for High Fidelity 3D Face Reconstruction
Gun-Hee Lee, Seong-Whan Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6100-6109
3D Morphable Model (3DMM) is a statistical model of facial shape and texture using a set of linear basis functions. Most of the recent 3D face reconstruction methods aim to embed the 3D morphable basis functions into Deep Convolutional Neural Network (DCNN). However, balancing the requirements of strong regularization for global shape and weak regularization for high level details is still ill-posed. To address this problem, we properly control generality and specificity in terms of regularization by harnessing the power of uncertainty. Additionally, we focus on the concept of nonlinearity and find out that Graph Convolutional Neural Network (Graph CNN) and Generative Adversarial Network (GAN) are effective in reconstructing high quality 3D shapes and textures respectively. In this paper, we propose to employ (i) an uncertainty-aware encoder that presents face features as distributions and (ii) a fully nonlinear decoder model combining Graph CNN with GAN. We demonstrate how our method builds excellent high quality results and outperforms previous state-of-the-art methods on 3D face reconstruction tasks for both constrained and in-the-wild images.
****

Object-Occluded Human Shape and Pose Estimation From a Single Color Image
Tianshu Zhang, Buzhen Huang, Yangang Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7376-7385
Occlusions between human and objects, especially for the activities of human-object interactions, are very common in practical applications. However, most of the existing approaches for 3D human shape and pose estimation require human bodies are well captured without occlusions or with minor self-occlusions. In this paper, we focus on the problem of directly estimating the object-occluded human shape and pose from single color images. Our key idea is to utilize a partial UV map to represent an object-occluded human body, and the full 3D human shape estimation is ultimately converted as an image inpainting problem. We propose a novel two-branch network architecture to train an end-to-end regressor via the latent feature supervision, which also includes a novel saliency map sub-net to extract the human information from object-occluded color images. To supervise the network training, we further build a novel dataset named as 3DOH50K. Several experiments are conducted to reveal the effectiveness of the proposed method. Experimental results demonstrate that the proposed method achieves the state-of-the-art comparing with previous methods. The dataset, codes are publicly available at https://www.yangangwang.com.
****

MAST: A Memory-Augmented Self-Supervised Tracker
Zihang Lai, Erika Lu, Weidi Xie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6479-6488
Recent interest in self-supervised dense tracking has yielded rapid progress, but performance still remains far from supervised methods. We propose a dense tracking model trained on videos without any annotations that surpasses previous self-supervised methods on existing benchmarks by a significant margin (+15%), and achieves performance comparable to supervised methods. In this paper, we first reassess the traditional choices used for self-supervised training and reconstruction loss by conducting thorough experiments that finally elucidate the optimal choices. Second, we further improve on existing methods by augmenting our architecture with a crucial memory component. Third, we benchmark on large-scale semi-supervised video object segmentation (aka. dense tracking), and propose a new metric: generalizability. Our first two contributions yield a self-supervised network that for the first time is competitive with supervised methods on standard evaluation metrics of dense tracking. When measuring generalizability, we show self-supervised approaches are actually superior to the majority of supervised methods. We believe this new generalizability metric can better capture the real-world use-cases for dense tracking, and will spur new interest in this research direction.
****

Wish You Were Here: Context-Aware Human Generation
Oran Gafni, Lior Wolf; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7840-7849
We present a novel method for inserting objects, specifically humans, into existing images, such that they blend in a photorealistic manner, while respecting the semantic context of the scene. Our method involves three subnetworks: the first generates the semantic map of the new person, given the pose of the other persons in the scene and an optional bounding box specification. The second network renders the pixels of the novel person and its blending mask, based on specifications in the form of multiple appearance components. A third network refines the generated face in order to match those of the target person. Our experiments present convincing high-resolution outputs in this novel and challenging application domain. In addition, the three networks are evaluated individually, demonstrating for example, state of the art results in pose transfer benchmarks.
****

Attention-Driven Cropping for Very High Resolution Facial Landmark Detection
Prashanth Chandran, Derek Bradley, Markus Gross, Thabo Beeler; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5861-5870
Facial landmark detection is a fundamental task for many consumer and high-end applications and is almost entirely solved by machine learning methods today. Existing datasets used to train such algorithms are primarily made up of only low resolution images, and current algorithms are limited to inputs of comparable quality and resolution as the training dataset. On the other hand, high resolution imagery is becoming increasingly more common as consumer cameras improve in quality every year. Therefore, there is need for algorithms that can leverage the rich information available in high resolution imagery. Naively attempting to reuse existing network architectures on high resolution imagery is prohibitive due to memory bottlenecks on GPUs. The only current solution is to downsample the images, sacrificing resolution and quality. Building on top of recent progress in attention-based networks, we present a novel, fully convolutional regional architecture that is specially designed for predicting landmarks on very high resolution facial images without downsampling. We demonstrate the flexibility of our architecture by training the proposed model with images of resolutions ranging from 256 x 256 to 4K. In addition to being the first method for facial landmark detection on high resolution images, our approach achieves superior performance over traditional (holistic) state-of-the-art architectures across ALL resolutions, leading to a general-purpose, extremely flexible, high quality landmark detector.
****

Contextual Residual Aggregation for Ultra High-Resolution Image Inpainting
Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, Zhan Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7508-7517
Recently data-driven image inpainting methods have made inspiring progress, impacting fundamental image editing tasks such as object removal and damaged image repairing. These methods are more effective than classic approaches, however, due to memory limitations they can only handle low-resolution inputs, typically smaller than 1K. Meanwhile, the resolution of photos captured with mobile devices increases up to 8K. Naive up-sampling of the low-resolution inpainted result can merely yield a large yet blurry result. Whereas, adding a high-frequency residual image onto the large blurry image can generate a sharp result, rich in details and textures. Motivated by this, we propose a Contextual Residual Aggregation (CRA) mechanism that can produce high-frequency residuals for missing contents by weighted aggregating residuals from contextual patches, thus only requiring a low-resolution prediction from the network. Since convolutional layers of the neural network only need to operate on low-resolution inputs and outputs, the cost of memory and computing power is thus well suppressed. Moreover, the need for high-resolution training datasets is alleviated. In our experiments, we train the proposed model on small images with resolutions 512 x 512 and perform inference on high-resolution images, achieving compelling inpainting quality. Our model can inpaint images as large as 8K with considerable hole sizes, which is intractable with previous learning-based approaches. We further elaborate on the light-weight design of the network architecture, achieving real-time performance on 2K images on a GTX 1080 Ti GPU. Codes are available at: https://github. com/Ascend-Huawei/Ascend-Canada/tree/ master/Models/Research_HiFIll_Model
****
StructEdit: Learning Structural Shape Variations
Kaichun Mo, Paul Guerrero, Li Yi, Hao Su, Peter Wonka, Niloy J. Mitra, Leonidas J. Guibas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8859-8868
Learning to encode differences in the geometry and (topological) structure of the shapes of ordinary objects is key to generating semantically plausible variations of a given shape, transferring edits from one shape to another, and for many other applications in 3D content creation. The common approach of encoding shapes as points in a high-dimensional latent feature space suggests treating shape differences as vectors in that space. Instead, we treat shape differences as primary objects in their own right and propose to encode them in their own latent space. In a setting where the shapes themselves are encoded in terms of fine-grained part hierarchies, we demonstrate that a separate encoding of shape deltas or differences provides a principled way to deal with inhomogeneities in the shape space due to different combinatorial part structures, while also allowing for compactness in the representation, as well as edit abstraction and transfer. Our approach is based on a conditional variational autoencoder for encoding and decoding shape deltas, conditioned on a source shape. We demonstrate the effectiveness and robustness of our approach in multiple shape modification and generation tasks, and provide comparison and ablation studies on the PartNet dataset, one of the largest publicly available 3D datasets.
****
Hierarchical Human Parsing With Typed Part-Relation Reasoning
Wenguan Wang, Hailong Zhu, Jifeng Dai, Yanwei Pang, Jianbing Shen, Ling Shao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8929-8939
Human parsing is for pixel-wise human semantic understanding. As human bodies are underlying hierarchically structured, how to model human structures is the central theme in this task. Focusing on this, we seek to simultaneously exploit the representational capacity of deep graph networks and the hierarchical human structures. In particular, we provide following two contributions. First, three kinds of part relations, i.e., decomposition, composition, and dependency, are, for the first time, completely and precisely described by three distinct relation n

etworks. This is in stark contrast to previous parsers, which only focus on a po rtion of the relations and adopt a type-agnostic relation modeling strategy. Mor e expressive relation information can be captured by explicitly imposing the par ameters in the relation networks to satisfy the specific characteristics of diff erent relations. Second, previous parsers largely ignore the need for an approxi mation algorithm over the loopy human hierarchy, while we instead address an ite rative reasoning process, by assimilating generic message-passing networks with their edge-typed, convolutional counterparts. With these efforts, our parser lay s the foundation for more sophisticated and flexible human relation patterns of reasoning. Comprehensive experiments on five datasets demonstrate that our parse r sets a new state-of-the-art on each.
****

High-Resolution Daytime Translation Without Domain Labels
Ivan Anokhin, Pavel Solovev, Denis Korzhenkov, Alexey Kharlamov, Taras Khakh ulin, Aleksei Silvestrov, Sergey Nikolenko, Victor Lempitsky, Gleb Sterkin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio n (CVPR), 2020, pp. 7488-7497
Modeling daytime changes in high resolution photographs, e.g., re-rendering the same scene under different illuminations typical for day, night, or dawn, is a c hallenging image manipulation task. We present the high-resolution daytime trans lation (HiDT) model for this task. HiDT combines a generative image-to-image mod el and a new upsampling scheme that allows to apply image translation at high re solution. The model demonstrates competitive results in terms of both commonly u sed GAN metrics and human evaluation. Importantly, this good performance comes a s a result of training on a dataset of still landscape images with no daytime la bels available.
****

Non-Adversarial Video Synthesis With Learned Priors
Abhishek Aich, Akash Gupta, Rameswar Panda, Rakib Hyder, M. Salman Asif, Am it K. Roy-Chowdhury; Proceedings of the IEEE/CVF Conference on Computer Vision a nd Pattern Recognition (CVPR), 2020, pp. 6090-6099
Most of the existing works in video synthesis focus on generating videos using a dversarial learning. Despite their success, these methods often require input re ference frame or fail to generate diverse videos from the given data distributio n, with little to no uniformity in the quality of videos that can be generated. Different from these methods, we focus on the problem of generating videos from latent noise vectors, without any reference input frames. To this end, we develo p a novel approach that jointly optimizes the input latent space, the weights of a recurrent neural network and a generator through non-adversarial learning. Op timizing for the input latent space along with the network weights allows us to generate videos in a controlled environment, i.e., we can faithfully generate al l videos the model has seen during the learning process as well as new unseen vi deos. Extensive experiments on three challenging and diverse datasets well demon strate that our proposed approach generates superior quality videos compared to the existing state-of-the-art methods.
****

Deep Homography Estimation for Dynamic Scenes
Hoang Le, Feng Liu, Shu Zhang, Aseem Agarwala; Proceedings of the IEEE/CVF Co nference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7652-7661
Homography estimation is an important step in many computer vision problems. Rec ently, deep neural network methods have shown to be favorable for this problem w hen compared to traditional methods. However, these new methods do not consider dynamic content in input images. They train neural networks with only image pair s that can be perfectly aligned using homographies. This paper investigates and discusses how to design and train a deep neural network that handles dynamic sce nes. We first collect a large video dataset with dynamic content. We then develo p a multi-scale neural network and show that when properly trained using our new dataset, this neural network can already handle dynamic scenes to some extent. To estimate a homography of a dynamic scene in a more principled way, we need to identify the dynamic content. Since dynamic content detection and homography es

timation are two tightly coupled tasks, we follow the multi-task learning princi ples and augment our multi-scale network such that it jointly estimates the dyna mics masks and homographies. Our experiments show that our method can robustly e stimate homography for challenging scenarios with dynamic scenes, blur artifacts , or lack of textures.
****

Where Does It End? - Reasoning About Hidden Surfaces by Object Intersection Cons traints
Michael Strecke, Jorg Stuckler; Proceedings of the IEEE/CVF Conference on Compu ter Vision and Pattern Recognition (CVPR), 2020, pp. 9592-9600
Dynamic scene understanding is an essential capability in robotics and VR/AR. In this paper we propose Co-Section, an optimization-based approach to 3D dynamic scene reconstruction, which infers hidden shape information from intersection co nstraints. An object-level dynamic SLAM frontend detects, segments, tracks and m aps dynamic objects in the scene. Our optimization backend completes the shapes using hull and intersection constraints between the objects. In experiments, we demonstrate our approach on real and synthetic dynamic scene datasets. We also a ssess the shape completion performance of our method quantitatively. To the best of our knowledge, our approach is the first method to incorporate such physical plausibility constraints on object intersections for shape completion of dynami c objects in an energy minimization framework.
****

Epipolar Transformers
Yihui He, Rui Yan, Katerina Fragkiadaki, Shoou-I Yu; Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7779 -7788
A common approach to localize 3D human joints in a synchronized and calibrated m ulti-view setup consists of two-steps: (1) apply a 2D detector separately on eac h view to localize joints in 2D, and (2) perform robust triangulation on 2D dete ctions from each view to acquire the 3D joint locations. However, in step 1, the 2D detector is limited to solving challenging cases which could potentially be better resolved in 3D, such as occlusions and oblique viewing angles, purely in 2D without leveraging any 3D information. Therefore, we propose the differentiab le "epipolar transformer", which enables the 2D detector to leverage 3D-aware fe atures to improve 2D pose estimation. The intuition is: given a 2D location p in the current view, we would like to first find its corresponding point p' in a n eighboring view, and then combine the features at p' with the features at p, thu s leading to a 3D-aware feature at p. Inspired by stereo matching, the epipolar transformer leverages epipolar constraints and feature matching to approximate t he features at p'. Experiments on InterHand and Human3.6M show that our approach has consistent improvements over the baselines. Specifically, in the condition where no external data is used, our Human3.6M model trained with ResNet-50 backb one and image size 256 x 256 outperforms state-of-the-art by 4.23 mm and achieve s MPJPE 26.9 mm.
****

Correlating Edge, Pose With Parsing
Ziwei Zhang, Chi Su, Liang Zheng, Xiaodong Xie; Proceedings of the IEEE/CVF C onference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8900-8909
According to existing studies, human body edge and pose are two beneficial facto rs to human parsing. The effectiveness of each of the high-level features (edge and pose) is confirmed through the concatenation of their features with the pars ing features. Driven by the insights, this paper studies how human semantic boun daries and keypoint locations can jointly improve human parsing. Compared with t he existing practice of feature concatenation, we find that uncovering the corre lation among the three factors is a superior way of leveraging the pivotal conte xtual cues provided by edges and poses. To capture such correlations, we propose a Correlation Parsing Machine (CorrPM) employing a heterogeneous non-local bloc k to discover the spatial affinity among feature maps from the edge, pose and pa rsing. The proposed CorrPM allows us to report new state-of-the-art accuracy on three human parsing datasets. Importantly, comparative studies confirm the advan

tages of feature correlation over the concatenation.
****
Relative Interior Rule in Block-Coordinate Descent
Tomas Werner, Daniel Prusa, Tomas Dlask; Proceedings of the IEEE/CVF Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7559-7567
It is well-known that for general convex optimization problems, block-coordinate descent can get stuck in poor local optima. Despite that, versions of this meth od known as convergent message passing are very successful to approximately solv e the dual LP relaxation of the MAP inference problem in graphical models. In at tempt to identify the reason why these methods often achieve good local minima, we argue that if in block-coordinate descent the set of minimizers over a variab le block has multiple elements, one should choose an element from the relative i nterior of this set. We show that this rule is not worse than any other rule for choosing block-minimizers. Based on this observation, we develop a theoretical framework for block-coordinate descent applied to general convex problems. We il lustrate this theory on convergent message-passing methods.
****
Controllable Person Image Synthesis With Attribute-Decomposed GAN
Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, Zhouhui Lian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2 020, pp. 5084-5093
This paper introduces the Attribute-Decomposed GAN, a novel generative model for controllable person image synthesis, which can produce realistic person images with desired human attributes (e.g., pose, head, upper clothes and pants) provid ed in various source inputs. The core idea of the proposed model is to embed hum an attributes into the latent space as independent codes and thus achieve flexib le and continuous control of attributes via mixing and interpolation operations in explicit style representations. Specifically, a new architecture consisting o f two encoding pathways with style block connections is proposed to decompose th e original hard mapping into multiple more accessible subtasks. In source pathwa y, we further extract component layouts with an off-the-shelf human parser and f eed them into a shared global texture encoder for decomposed latent codes. This strategy allows for the synthesis of more realistic output images and automatic separation of un-annotated attributes. Experimental results demonstrate the prop osed method's superiority over the state of the art in pose transfer and its eff ectiveness in the brand-new task of component attribute transfer.
****
Unpaired Portrait Drawing Generation via Asymmetric Cycle Mapping
Ran Yi, Yong-Jin Liu, Yu-Kun Lai, Paul L. Rosin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8217-822 5
Portrait drawing is a common form of art with high abstraction and expressivenes s. Due to its unique characteristics, existing methods achieve decent results on ly with paired training data, which is costly and time-consuming to obtain.In th is paper, we address the problem of automatic transfer from face photos to portr ait drawings with unpaired training data. We observe that due to the significant imbalance of information richness between photos and drawings, existing unpaire d transfer methods such as CycleGAN tends to embed invisible reconstruction info rmation indiscriminately in the whole drawings, leading to important facial feat ures partially missing in drawings. To address this problem, we propose a novel asymmetric cycle mapping that enforces the reconstruction information to be visi ble (by a truncation loss) and only embedded in selective facial regions (by a r elaxed forward cycle-consistency loss). Along with localized discriminators for the eyes, nose and lips, our method well preserves all important facial features in the generated portrait drawings. By introducing a style classifier and takin g the style vector into account, our method can learn to generate portrait drawi ngs in multiple styles using a single network. Extensive experiments show that o ur model outperforms state-of-the-art methods.
****
Advancing High Fidelity Identity Swapping for Forgery Detection

Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, Fang Wen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5074-5083

In this work, we study various existing benchmarks for deepfake detection researches. In particular, we examine a novel two-stage face swapping algorithm, called FaceShifter, for high fidelity and occlusion aware face swapping. Unlike many existing face swapping works that leverage only limited information from the target image when synthesizing the swapped face, FaceShifter generates the swapped face with high-fidelity by exploiting and integrating the target attributes thoroughly and adaptively. FaceShifter can handle facial occlusions with a second synthesis stage consisting of a Heuristic Error Acknowledging Refinement Network (HEAR-Net), which is trained to recover anomaly regions in a self-supervised way without any manual annotations. Experiments show that existing deepfake detection algorithm performs poorly with FaceShifter, since it achieves advantageous quality over all existing benchmarks. However, our newly developed Face X-Ray method can reliably detect forged images created by FaceShifter.

****

BachGAN: High-Resolution Image Synthesis From Salient Object Layout

Yandong Li, Yu Cheng, Zhe Gan, Licheng Yu, Liqiang Wang, Jingjing Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8365-8374

We propose a new task towards more practical applications for image generation - high-quality image synthesis from salient object layout. This new setting requires users to provide only the layout of salient objects (i.e., foreground bounding boxes and categories) and lets the model complete the drawing with an invented background and a matching foreground. Two main challenges spring from this new task: (i) how to generate fine-grained details and realistic textures without segmentation map input; and (ii) how to create and weave a background into standalone objects in a seamless way. To tackle this, we propose Background Hallucination Generative Adversarial Network (BachGAN), which leverages a background retrieval module to first select a set of segmentation maps from a large candidate pool, then encodes these candidate layouts via a background fusion module to hallucinate a suitable background for the given objects. By generating the hallucinated background representation dynamically, our model can synthesize high-resolution images with both photo-realistic foreground and integral background. Experiments on Cityscapes and ADE20K datasets demonstrate the advantage of BachGAN over existing approaches, measured on both visual fidelity of generated images and visual alignment between output images and input layouts.

****

SER-FIQ: Unsupervised Estimation of Face Image Quality Based on Stochastic Embedding Robustness

Philipp Terhorst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, Arjan Kuijper; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5651-5660

Face image quality is an important factor to enable high-performance face recognition systems. Face quality assessment aims at estimating the suitability of a face image for the purpose of recognition. Previous work proposed supervised solutions that require artificially or human labelled quality values. However, both labelling mechanisms are error prone as they do not rely on a clear definition of quality and may not know the best characteristics for the utilized face recognition system. Avoiding the use of inaccurate quality labels, we proposed a novel concept to measure face quality based on an arbitrary face recognition model. By determining the embedding variations generated from random subnetworks of a face model, the robustness of a sample representation and thus, its quality is estimated. The experiments are conducted in a cross-database evaluation setting on three publicly available databases. We compare our proposed solution on two face embeddings against six state-of-the-art approaches from academia and industry. The results show that our unsupervised solution outperforms all other approaches in the majority of the investigated scenarios. In contrast to previous works, the proposed solution shows a stable performance over all scenarios. Utilizing th

e deployed face recognition model for our face quality assessment methodology av oids the training phase completely and further outperforms all baseline approach es by a large margin. Our solution can be easily integrated into current face re cognition systems, and can be modified to other tasks beyond face recognition.
****

Globally Optimal Contrast Maximisation for Event-Based Motion Estimation
Daqi Liu, Alvaro Parra, Tat-Jun Chin; Proceedings of the IEEE/CVF Conference o n Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6349-6358
Contrast maximisation estimates the motion captured in an event stream by maximi sing the sharpness of the motion-compensated event image. To carry out contrast maximisation, many previous works employ iterative optimisation algorithms, such as conjugate gradient, which require good initialisation to avoid converging to bad local minima. To alleviate this weakness, we propose a new globally optimal event-based motion estimation algorithm. Based on branch-and-bound (BnB), our m ethod solves rotational (3DoF) motion estimation on event streams, which support s practical applications such as video stabilisation and attitude estimation. Un derpinning our method are novel bounding functions for contrast maximisation, wh ose theoretical validity is rigorously established. We show concrete examples fr om public datasets where globally optimal solutions are vital to the success of contrast maximisation. Despite its exact nature, our algorithm is currently able to process a 50,000-event input in approx 300 seconds (a locally optimal solver takes approx 30 seconds on the same input), and has the potential to be further speeded-up using GPUs.
****

Towards High-Fidelity 3D Face Reconstruction From In-the-Wild Images Using Graph Convolutional Networks
Jiangke Lin, Yi Yuan, Tianjia Shao, Kun Zhou; Proceedings of the IEEE/CVF Con ference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5891-5900
3D Morphable Model (3DMM) based methods have achieved great success in recoverin g 3D face shapes from single-view images. However, the facial textures recovered by such methods lack the fidelity as exhibited in the input images. Recent work s demonstrate high-quality facial texture recovering with generative networks tr ained from a large-scale database of high-resolution UV maps of face textures, w hich is hard to prepare and not publicly available. In this paper, we introduce a method to reconstruct 3D facial shapes with high-fidelity textures from single -view images in the wild, without the need to capture a large-scale face texture database. The main idea is to refine the initial texture generated by a 3DMM ba sed method with facial details from the input image. To this end, we propose to use graph convolutional networks to reconstruct the detailed colors for the mesh vertices instead of reconstructing the UV map. Experiments show that our method can generate high-quality results and outperforms state-of-the-art methods in b oth qualitative and quantitative comparisons.
****

PolyTransform: Deep Polygon Transformer for Instance Segmentation
Justin Liang, Namdar Homayounfar, Wei-Chiu Ma, Yuwen Xiong, Rui Hu, Raquel Urtasun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern R ecognition (CVPR), 2020, pp. 9131-9140
In this paper, we propose PolyTransform, a novel instance segmentation algorithm that produces precise, geometry-preserving masks by combining the strengths of prevailing segmentation approaches and modern polygon-based methods. In particul ar, we first exploit a segmentation network to generate instance masks. We then convert the masks into a set of polygons that are then fed to a deforming networ k that transforms the polygons such that they better fit the object boundaries. Our experiments on the challenging Cityscapes dataset show that our PolyTransfor m significantly improves the performance of the backbone instance segmentation n etwork and ranks 1st on the Cityscapes test-set leaderboard. We also show impres sive gains in the interactive annotation setting.
****

Towards Fairness in Visual Recognition: Effective Strategies for Bias Mitigation
Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair,

Kenji Hata, Olga Russakovsky; Proceedings of the IEEE/CVF Conference on Compu ter Vision and Pattern Recognition (CVPR), 2020, pp. 8919-8928

Computer vision models learn to perform a task by capturing relevant statistics from training data. It has been shown that models learn spurious age, gender, an d race correlations when trained for seemingly unrelated tasks like activity rec ognition or image captioning. Various mitigation techniques have been presented to prevent models from utilizing or learning such biases. However, there has bee n little systematic comparison between these techniques. We design a simple but surprisingly effective visual recognition benchmark for studying bias mitigation . Using this benchmark, we provide a thorough analysis of a wide range of techni ques. We highlight the shortcomings of popular adversarial training approaches f or bias mitigation, propose a simple but similarly effective alternative to the inference-time Reducing Bias Amplification method of Zhao et al., and design a d omain-independent training technique that outperforms all other methods. Finally , we validate our findings on the attribute classification task in the CelebA da taset, where attribute presence is known to be correlated with the gender of peo ple in the image, and demonstrate that the proposed technique is effective at mi tigating real-world gender bias.

****

RDCFace: Radial Distortion Correction for Face Recognition

He Zhao, Xianghua Ying, Yongjie Shi, Xin Tong, Jingsi Wen, Hongbin Zha; Pro ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7721-7730

The effects of radial lens distortion often appear in wide-angle cameras of surv eillance and safeguard systems, which may severely degrade performances of previ ous face recognition algorithms. Traditional methods for radial lens distortion correction usually employ line features in scenarios that are not suitable for f ace images. In this paper, we propose a distortion-invariant face recognition sy stem called RDCFace, which directly and only utilize the distorted images of fac es, to alleviate the effects of radial lens distortion. RDCFace is an end-to-end trainable cascade network, which can learn rectification and alignment parame ters to achieve a better face recognition performance without requiring supervisio n of facial landmarks and distortion parameters. We design sequential spatial tr ansformer layers to optimize the correction, alignment, and recognition modules jointly. The feasibility of our method comes from implicitly using the statistic s of the layout of face features learned from the large-scale face data. Extensi ve experiments indicate that our method is distortion robust and gains significa nt improvements on LFW, YTF, CFP, and RadialFace, a real distorted face benchmar k compared with state-of-the-art methods.

****

Learning Dynamic Routing for Semantic Segmentation

Yanwei Li, Lin Song, Yukang Chen, Zeming Li, Xiangyu Zhang, Xingang Wang, Jian Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8553-8562

Recently, numerous handcrafted and searched networks have been applied for seman tic segmentation. However, previous works intend to handle inputs with various s cales in pre-defined static architectures, such as FCN, U-Net, and DeepLab serie s. This paper studies a conceptually new method to alleviate the scale variance in semantic representation, named dynamic routing. The proposed framework genera tes data-dependent routes, adapting to the scale distribution of each image. To this end, a differentiable gating function, called soft conditional gate, is pro posed to select scale transform paths on the fly. In addition, the computational cost can be further reduced in an end-to-end manner by giving budget constraint s to the gating function. We further relax the network level routing space to su pport multi-path propagations and skip-connections in each forward, bringing sub stantial network capacity. To demonstrate the superiority of the dynamic propert y, we compare with several static architectures, which can be modeled as special cases in the routing space. Extensive experiments are conducted on Cityscapes a nd PASCAL VOC 2012 to illustrate the effectiveness of the dynamic framework. Cod e is available at https://github.com/yanwei-li/DynamicRouting.

****
GNN3DMOT: Graph Neural Network for 3D Multi-Object Tracking With 2D-3D Multi-Feature Learning
Xinshuo Weng, Yongxin Wang, Yunze Man, Kris M. Kitani; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6499-6508
3D Multi-object tracking (MOT) is crucial to autonomous systems. Recent work uses a standard tracking-by-detection pipeline, where feature extraction is first performed independently for each object in order to compute an affinity matrix. Then the affinity matrix is passed to the Hungarian algorithm for data association. A key process of this standard pipeline is to learn discriminative features for different objects in order to reduce confusion during data association. In this work, we propose two techniques to improve the discriminative feature learning for MOT: (1) instead of obtaining features for each object independently, we propose a novel feature interaction mechanism by introducing the Graph Neural Network. As a result, the feature of one object is informed of the features of other objects so that the object feature can lean towards the object with similar feature (i.e., object probably with a same ID) and deviate from objects with dissimilar features (i.e., object probably with different IDs), leading to a more discriminative feature for each object; (2) instead of obtaining the feature from either 2D or 3D space in prior work, we propose a novel joint feature extractor to learn appearance and motion features from 2D and 3D space simultaneously. As features from different modalities often have complementary information, the joint feature can be more discriminate than feature from each individual modality. To ensure that the joint feature extractor does not heavily rely on one modality, we also propose an ensemble training paradigm. Through extensive evaluation, our proposed method achieves state-of-the-art performance on KITTI and nuScenes 3D MOT benchmarks. Our code will be made available at https://github.com/xinshuoweng/GNN3DMOT
****
Searching Central Difference Convolutional Networks for Face Anti-Spoofing
Zitong Yu, Chenxu Zhao, Zezheng Wang, Yunxiao Qin, Zhuo Su, Xiaobai Li, Feng Zhou, Guoying Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5295-5305
Face anti-spoofing (FAS) plays a vital role in face recognition systems. Most state-of-the-art FAS methods 1) rely on stacked convolutions and expert-designed network, which is weak in describing detailed fine-grained information and easily being ineffective when the environment varies (e.g., different illumination), and 2) prefer to use long sequence as input to extract dynamic features, making them difficult to deploy into scenarios which need quick response. Here we propose a novel frame level FAS method based on Central Difference Convolution (CDC), which is able to capture intrinsic detailed patterns via aggregating both intensity and gradient information. A network built with CDC, called the Central Difference Convolutional Network (CDCN), is able to provide more robust modeling capacity than its counterpart built with vanilla convolution. Furthermore, over a specifically designed CDC search space, Neural Architecture Search (NAS) is utilized to discover a more powerful network structure (CDCN++), which can be assembled with Multiscale Attention Fusion Module (MAFM) for further boosting performance. Comprehensive experiments are performed on six benchmark datasets to show that 1) the proposed method not only achieves superior performance on intra-dataset testing (especially 0.2% ACER in Protocol-1 of OULU-NPU dataset), 2) it also generalizes well on cross-dataset testing (particularly 6.5% HTER from CASIA-MFSD to Replay-Attack datasets). The codes are available at https://github.com/ZitongYu/CDCN.
****
PREDICT & CLUSTER: Unsupervised Skeleton Based Action Recognition
Kun Su, Xiulong Liu, Eli Shlizerman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9631-9640
We propose a novel system for unsupervised skeleton-based action recognition. Given inputs of body-keypoints sequences obtained during various movements, our sy

stem associates the sequences with actions. Our system is based on an encoder-de coder recurrent neural network, where the encoder learns a separable feature rep resentation within its hidden states formed by training the model to perform the prediction task. We show that according to such unsupervised training, the deco der and the encoder self-organize their hidden states into a feature space which clusters similar movements into the same cluster and distinct movements into di stant clusters. Current state-of-the-art methods for action recognition are stro ngly supervised, i.e., rely on providing labels for training. Unsupervised metho ds have been proposed, however, they require camera and depth inputs (RGB+D) at each time step. In contrast, our system is fully unsupervised, does not require action labels at any stage and can operate with body-keypoints input only. Furth ermore, the method can perform on various dimensions of body-keypoints (2D or 3D) and can include additional cues describing movements. We evaluate our system o n three action recognition benchmarks with different numbers of actions and exam ples. Our results outperform prior unsupervised skeleton-based methods, unsuperv ised RGB+D based methods on cross-view tests and while being unsupervised have s imilar performance to supervised skeleton-based action recognition.
****

RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild
Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, Stefanos Zafeiriou ; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit ion (CVPR), 2020, pp. 5203-5212
Though tremendous strides have been made in uncontrolled face detection, accurat e and efficient 2D face alignment and 3D face reconstruction in-the-wild remain an open challenge. In this paper, we present a novel single-shot, multi-level fa ce localisation method, named RetinaFace, which unifies face box prediction, 2D facial landmark localisation and 3D vertices regression under one common target: point regression on the image plane. To fill the data gap, we manually annotate d five facial landmarks on the WIDER FACE dataset and employed a semi-automatic annotation pipeline to generate 3D vertices for face images from the WIDER FACE, AFLW and FDDB datasets. Based on extra annotations, we propose a mutually benef icial regression target for 3D face reconstruction, that is predicting 3D vertic es projected on the image plane constrained by a common 3D topology. The propose d 3D face reconstruction branch can be easily incorporated, without any optimisa tion difficulty, in parallel with the existing box and 2D landmark regression br anches during joint training. Extensive experimental results show that RetinaFac e can simultaneously achieve stable face detection, accurate 2D face alignment a nd robust 3D face reconstruction while being efficient through single-shot infer ence.
****

Monocular Real-Time Hand Shape and Motion Capture Using Multi-Modal Data
Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobal t, Feng Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Patte rn Recognition (CVPR), 2020, pp. 5346-5355
We present a novel method for monocular hand shape and pose estimation at unprec edented runtime performance of 100fps and at state-of-the-art accuracy. This is enabled by a new learning based architecture designed such that it can make use of all the sources of available hand training data: image data with either 2D or 3D annotations, as well as stand-alone 3D animations without corresponding imag e data. It features a 3D hand joint detection module and an inverse kinematics m odule which regresses not only 3D joint positions but also maps them to joint ro tations in a single feed-forward pass. This output makes the method more directl y usable for applications in computer vision and graphics compared to only regre ssing 3D joint positions. We demonstrate that our architectural design leads to a significant quantitative and qualitative improvement over the state of the art on several challenging benchmarks. We will make our code publicly available for future research.
****

Mitigating Bias in Face Recognition Using Skewness-Aware Reinforcement Learning
Mei Wang, Weihong Deng; Proceedings of the IEEE/CVF Conference on Computer Visi

on and Pattern Recognition (CVPR), 2020, pp. 9322-9331
Racial equality is an important theme of international human rights law, but it has been largely obscured when the overall face recognition accuracy is pursued blindly. More facts indicate racial bias indeed degrades the fairness of recognition system and the error rates on non-Caucasians are usually much higher than Caucasians. To encourage fairness, we introduce the idea of adaptive margin to learn balanced performance for different races based on large margin losses. A reinforcement learning based race balance network (RL-RBN) is proposed. We formulate the process of finding the optimal margins for non-Caucasians as a Markov decision process and employ deep Q-learning to learn policies for an agent to select appropriate margin by approximating the Q-value function. Guided by the agent, the skewness of feature scatter between races can be reduced. Besides, we provide two ethnicity aware training datasets, called BUPT-Globalface and BUPT-Balanceface dataset, which can be utilized to study racial bias from both data and algorithm aspects. Extensive experiments on RFW database show that RL-RBN successfully mitigates racial bias and learns more balanced performance.
****

Single Image Reflection Removal With Physically-Based Training Images
Soomin Kim, Yuchi Huo, Sung-Eui Yoon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5164-5173
Recently, deep learning-based single image reflection separation methods have been exploited widely. To benefit the learning approach, a large number of training image pairs (i.e., with and without reflections) were synthesized in various ways, yet they are away from a physically-based direction. In this paper, physically based rendering is used for faithfully synthesizing the required training images, and a corresponding network structure and loss term are proposed. We utilize existing RGBD/RGB images to estimate meshes, then physically simulate the light transportation between meshes, glass, and lens with path tracing to synthesize training data, which successfully reproduce the spatially variant anisotropic visual effect of glass reflection. For guiding the separation better, we additionally consider a module, backtrack network (BT-net) for backtracking the reflections, which removes complicated ghosting, attenuation, blurred and defocused effect of glass/lens. This enables obtaining a priori information before having the distortion. The proposed method considering additional a priori information with physically simulated training data is validated with various real reflection images and shows visually pleasant and numerical advantages compared with state-of-the-art techniques.
****

Disentangled Image Generation Through Structured Noise Injection
Yazeed Alharbi, Peter Wonka; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5134-5142
We explore different design choices for injecting noise into generative adversarial networks (GANs) with the goal of disentangling the latent space. Instead of traditional approaches, we propose feeding multiple noise codes through separate fully-connected layers respectively. The aim is restricting the influence of each noise code to specific parts of the generated image. We show that disentanglement in the first layer of the generator network leads to disentanglement in the generated image. Through a grid-based structure, we achieve several aspects of disentanglement without complicating the network architecture and without requiring labels. We achieve spatial disentanglement, scale-space disentanglement, and disentanglement of the foreground object from the background style allowing fine-grained control over the generated images. Examples include changing facial expressions in face images, changing beak length in bird images, and changing car dimensions in car images. This empirically leads to better disentanglement scores than state-of-the-art methods on the FFHQ dataset.
****

Deep 3D Capture: Geometry and Reflectance From Sparse Multi-View Images
Sai Bi, Zexiang Xu, Kalyan Sunkavalli, David Kriegman, Ravi Ramamoorthi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5960-5969

We introduce a novel learning-based method to reconstruct the high-quality geometry and complex, spatially-varying BRDF of an arbitrary object from a sparse set of only six images captured by wide-baseline cameras under collocated point lighting. We first estimate per-view depth maps using a deep multi-view stereo network; these depth maps are used to coarsely align the different views. We propose a novel multi-view reflectance estimation network architecture that is trained to pool features from these coarsely aligned images and predict per-view spatially-varying diffuse albedo, surface normals, specular roughness and specular albedo. We do this by jointly optimizing the latent space of our multi-view reflectance network to minimize the photometric error between images rendered with our predictions and the input images. While previous state-of-the-art methods fail on such sparse acquisition setups, we demonstrate, via extensive experiments on synthetic and real data, that our method produces high-quality reconstructions that can be used to render photorealistic images.

****

Multi-Scale Fusion Subspace Clustering Using Similarity Constraint

Zhiyuan Dang, Cheng Deng, Xu Yang, Heng Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6658-6667

Classical subspace clustering methods often assume that the raw form data lie in a union of the low-dimension linear subspace. This assumption is too strict in practice, which largely limits the generalization of subspace clustering. To tackle this issue, deep subspace clustering (DSC) networks based on deep autoencoder (DAE) have been proposed, which non-linearly map the raw form data into a latent space well-adapted to subspace clustering. However, existing DSC models ignore the important multi-scale information embedded in DAE, thus abandon the much more useful deep features, leading their suboptimal clustering results. In this paper, we propose the Multi-Scale Fusion Subspace Clustering Using Similarity Constraint (SC-MSFSC) network, which learns a more discriminative self-expression coefficient matrix by a novel multi-scale fusion module. More importantly, it introduces a similarity constraint module to guide the fused self-expression coefficient matrix in training. Specifically, the multi-scale fusion module is framed to generate the self-expression coefficient matrix of each convolutional layer in DAE and then fuses them with the convolutional kernel. In addition, the similarity constraint module is to supervise the fused self-expression coefficient matrix by the designed similarity matrix. Extensive experimental results on four benchmark datasets demonstrate the superiority of our new model against state-of-the-art methods.

****

GroupFace: Learning Latent Groups and Constructing Group-Based Representations for Face Recognition

Yonghyun Kim, Wonpyo Park, Myung-Cheol Roh, Jongju Shin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5621-5630

In the field of face recognition, a model learns to distinguish millions of face images with fewer dimensional embedding features, and such vast information may not be properly encoded in the conventional model with a single branch. We propose a novel face-recognition-specialized architecture called GroupFace that utilizes multiple group-aware representations, simultaneously, to improve the quality of the embedding feature. The proposed method provides self-distributed labels that balance the number of samples belonging to each group without additional human annotations, and learns the group-aware representations that can narrow down the search space of the target identity. We prove the effectiveness of the proposed method by showing extensive ablation studies and visualizations. All the components of the proposed method can be trained in an end-to-end manner with a marginal increase of computational complexity. Finally, the proposed method achieves the state-of-the-art results with significant improvements in 1:1 face verification and 1:N face identification tasks on the following public datasets: LFW, YTF, CALFW, CPLFW, CFP, AgeDB-30, MegaFace, IJB-B and IJB-C.

****

Learning to Optimize Non-Rigid Tracking

Yang Li, Aljaz Bozic, Tianwei Zhang, Yanli Ji, Tatsuya Harada, Matthias Niessner; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4910-4918
One of the widespread solutions for non-rigid tracking has a nested-loop structure: with Gauss-Newton to minimize a tracking objective in the outer loop, and Preconditioned Conjugate Gradient (PCG) to solve a sparse linear system in the inner loop. In this paper, we employ learnable optimizations to improve tracking robustness and speed up solver convergence. First, we upgrade the tracking objective by integrating an alignment data term on deep features which are learned end-to-end through CNN. The new tracking objective can capture the global deformation which helps Gauss-Newton to jump over local minimum, leading to robust tracking on large non-rigid motions. Second, we bridge the gap between the preconditioning technique and learning method by introducing a ConditionNet which is trained to generate a preconditioner such that PCG can converge within a small number of steps. Experimental results indicate that the proposed learning method converges faster than the original PCG by a large margin.
****

Weakly Supervised Discriminative Feature Learning With State Information for Person Identification
Hong-Xing Yu, Wei-Shi Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5528-5538
Unsupervised learning of identity-discriminative visual feature is appealing in real-world tasks where manual labelling is costly. However, the images of an identity can be visually discrepant when images are taken under different states, e.g. different camera views and poses. This visual discrepancy leads to great difficulty in unsupervised discriminative learning. Fortunately, in real-world tasks we could often know the states without human annotation, e.g. we can easily have the camera view labels in person re-identification and facial pose labels in face recognition. In this work we propose utilizing the state information as weak supervision to address the visual discrepancy caused by different states. We formulate a simple pseudo label model and utilize the state information in an attempt to refine the assigned pseudo labels by the weakly supervised decision boundary rectification and weakly supervised feature drift regularization. We evaluate our model on unsupervised person re-identification and pose-invariant face recognition. Despite the simplicity of our method, it could outperform the state-of-the-art results on Duke-reID, MultiPIE and CFP datasets with a standard ResNet-50 backbone. We also find our model could perform comparably with the standard supervised fine-tuning results on the three datasets. Code is available at https://github.com/KovenYu/state-information.
****

An Internal Covariate Shift Bounding Algorithm for Deep Neural Networks by Unitizing Layers' Outputs
You Huang, Yuanlong Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8465-8473
Batch Normalization (BN) techniques have been proposed to reduce the so-called Internal Covariate Shift (ICS) by attempting to keep the distributions of layer outputs unchanged. Experiments have shown their effectiveness on training deep neural networks. However, since only the first two moments are controlled in these BN techniques, it seems that a weak constraint is imposed on layer distributions and furthermore whether such constraint can reduce ICS is unknown. Thus this paper proposes a measure for ICS by using the Earth Mover (EM) distance and then derives the upper and lower bounds for the measure to provide a theoretical analysis of BN. The upper bound has shown that BN techniques can control ICS only for the outputs with low dimensions and small noise whereas their control is not effective in other cases. This paper also proves that such control is just a bounding of ICS rather than a reduction of ICS. Meanwhile, the analysis shows that the high-order moments and noise, which BN cannot control, have great impact on the lower bound. Based on such analysis, this paper furthermore proposes an algorithm that unitizes the outputs with an adjustable parameter to further bound ICS in order to cope with the problems of BN. The upper bound for the proposed unit

ization is noise-free and only dominated by the parameter. Thus, the parameter can be trained to tune the bound and further to control ICS. Besides, the unitization is embedded into the framework of BN to reduce the information loss. The experiments show that this proposed algorithm outperforms existing BN techniques on CIFAR-10, CIFAR-100 and ImageNet datasets.
****

MixNMatch: Multifactor Disentanglement and Encoding for Conditional Image Generation
Yuheng Li, Krishna Kumar Singh, Utkarsh Ojha, Yong Jae Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8039-8048
We present MixNMatch, a conditional generative model that learns to disentangle and encode background, object pose, shape, and texture from real images with minimal supervision, for mix-and-match image generation. We build upon FineGAN, an unconditional generative model, to learn the desired disentanglement and image generator, and leverage adversarial joint image-code distribution matching to learn the latent factor encoders. MixNMatch requires bounding boxes during training to model background, but requires no other supervision. Through extensive experiments, we demonstrate MixNMatch's ability to accurately disentangle, encode, and combine multiple factors for mix-and-match image generation, including sketch2color, cartoon2img, and img2gif applications. Our code/models/demo can be found at https://github.com/Yuheng-Li/MixNMatch
****

Parsing-Based View-Aware Embedding Network for Vehicle Re-Identification
Dechao Meng, Liang Li, Xuejing Liu, Yadong Li, Shijie Yang, Zheng-Jun Zha, Xingyu Gao, Shuhui Wang, Qingming Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7103-7112
Vehicle Re-Identification is to find images of the same vehicle from various views in the cross-camera scenario. The main challenges of this task are the large intra-instance distance caused by different views and the subtle inter-instance discrepancy caused by similar vehicles. In this paper, we propose a parsing-based view-aware embedding network (PVEN) to achieve the view-aware feature alignment and enhancement for vehicle ReID. First, we introduce a parsing network to parse a vehicle into four different views and then align the features by mask average pooling. Such alignment provides a fine-grained representation of the vehicle. Second, in order to enhance the view-aware features, we design a common-visible attention to focus on the common visible views, which not only shortens the distance among intra-instances, but also enlarges the discrepancy of inter-instances. The PVEN helps capture the stable discriminative information of vehicle under different views. The experiments conducted on three datasets show that our model outperforms state-of-the-art methods by a large margin.
****

PF-Net: Point Fractal Network for 3D Point Cloud Completion
Zitian Huang, Yikuan Yu, Jiawen Xu, Feng Ni, Xinyi Le; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7662-7670
In this paper, we propose a Point Fractal Network (PF-Net), a novel learning-based approach for precise and high-fidelity point cloud completion. Unlike existing point cloud completion networks, which generate the overall shape of the point cloud from the incomplete point cloud and always change existing points and encounter noise and geometrical loss, PF-Net preserves the spatial arrangements of the incomplete point cloud and can figure out the detailed geometrical structure of the missing region(s) in the prediction. To succeed at this task, PF-Net estimates the missing point cloud hierarchically by utilizing a feature-points-based multi-scale generating network. Further, we add up multi-stage completion loss and adversarial loss to generate more realistic missing region(s). The adversarial loss can better tackle multiple modes in the prediction. Our experiments demonstrate the effectiveness of our method for several challenging point cloud completion tasks.
****

Stochastic Classifiers for Unsupervised Domain Adaptation
Zhihe Lu, Yongxin Yang, Xiatian Zhu, Cong Liu, Yi-Zhe Song, Tao Xiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9111-9120
A common strategy adopted by existing state-of-the-art unsupervised domain adaptation (UDA) methods is to employ two classifiers to identify the misaligned local regions between source and target domain. Following the 'wisdom of the crowd' principle, one has to ask: why stop at two? Indeed, we find that using more classifiers leads to better performance, but also introduces more model parameters, therefore risking overfitting. In this paper, we introduce a novel method called STochastic clAssifieRs (STAR) for addressing this problem. Instead of representing one classifier as a weight vector, STAR models it as a Gaussian distribution with its variance representing the inter-classifier discrepancy. With STAR, we can now sample an arbitrary number of classifiers from the distribution, whilst keeping the model size the same as having two classifiers. Extensive experiments demonstrate that a variety of existing UDA methods can greatly benefit from STAR and achieve the state-of-the-art performance on both image classification and semantic segmentation tasks.
****

CIAGAN: Conditional Identity Anonymization Generative Adversarial Networks
Maxim Maximov, Ismail Elezi, Laura Leal-Taixe; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5447-5456
The unprecedented increase in the usage of computer vision technology in society goes hand in hand with an increased concern in data privacy. In many real-world scenarios like people tracking or action recognition, it is important to be able to process the data while taking careful consideration in protecting people's identity. We propose and develop CIAGAN, a model for image and video anonymization based on conditional generative adversarial networks. Our model is able to remove the identifying characteristics of faces and bodies while producing high-quality images and videos that can be used for any computer vision task, such as detection or tracking. Unlike previous methods, we have full control over the de-identification (anonymization) procedure, ensuring both anonymization as well as diversity. We compare our method to several baselines and achieve state-of-the-art results. To facilitate further research, we make available the code and the models at https://github.com/dvl-tum/ciagan.
****

Hierarchically Robust Representation Learning
Qi Qian, Juhua Hu, Hao Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7336-7344
With the tremendous success of deep learning in visual tasks, the representations extracted from intermediate layers of learned models, that is, deep features, attract much attention of researchers. Previous empirical analysis shows that those features can contain appropriate semantic information. Therefore, with a model trained on a large-scale benchmark data set (e.g., ImageNet), the extracted features can work well on other tasks. In this work, we investigate this phenomenon and demonstrate that deep features can be suboptimal due to the fact that they are learned by minimizing the empirical risk. When the data distribution of the target task is different from that of the benchmark data set, the performance of deep features can degrade. Hence, we propose a hierarchically robust optimization method to learn more generic features. Considering the example-level and concept-level robustness simultaneously, we formulate the problem as a distributionally robust optimization problem with Wasserstein ambiguity set constraints, and an efficient algorithm with the conventional training pipeline is proposed. Experiments on benchmark data sets demonstrate the effectiveness of the robust deep representations.
****

Towards Robust Image Classification Using Sequential Attention Models
Daniel Zoran, Mike Chrzanowski, Po-Sen Huang, Sven Gowal, Alex Mott, Pushmeet Kohli; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9483-9492

In this paper we propose to augment a modern neural-network architecture with an attention model inspired by human perception. Specifically, we adversarially train and analyze a neural model incorporating a human inspired, visual attention component that is guided by a recurrent top-down sequential process. Our experimental evaluation uncovers several notable findings about the robustness and behavior of this new model. First, introducing attention to the model significantly improves adversarial robustness resulting in state-of-the-art ImageNet accuracies under a wide range of random targeted attack strengths. Second, we show that by varying the number of attention steps (glances/fixations) for which the model is unrolled, we are able to make its defense capabilities stronger, even in light of stronger attacks --- resulting in a "computational race" between the attacker and the defender. Finally, we show that some of the adversarial examples generated by attacking our model are quite different from conventional adversarial examples --- they contain global, salient and spatially coherent structures coming from the target class that would be recognizable even to a human, and work by distracting the attention of the model away from the main object in the original image.

****

A Morphable Face Albedo Model

William A. P. Smith, Alassane Seck, Hannah Dee, Bernard Tiddeman, Joshua B. Tenenbaum, Bernhard Egger; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5011-5020

In this paper, we bring together two divergent strands of research: photometric face capture and statistical 3D face appearance modelling. We propose a novel lightstage capture and processing pipeline for acquiring ear-to-ear, truly intrinsic diffuse and specular albedo maps that fully factor out the effects of illumination, camera and geometry. Using this pipeline, we capture a dataset of 50 scans and combine them with the only existing publicly available albedo dataset (3DRFE) of 23 scans. This allows us to build the first morphable face albedo model. We believe this is the first statistical analysis of the variability of facial specular albedo maps. This model can be used as a plug in replacement for the texture model of the Basel Face Model and we make our new albedo model publicly available. We ensure careful spectral calibration such that our model is built in a linear sRGB space, suitable for inverse rendering of images taken by typical cameras. We demonstrate our model in a state of the art analysis-by-synthesis 3DMM fitting pipeline, are the first to integrate specular map estimation and outperform the Basel Face Model in albedo reconstruction.

****

Fast Video Object Segmentation With Temporal Aggregation Network and Dynamic Template Matching

Xuhua Huang, Jiarui Xu, Yu-Wing Tai, Chi-Keung Tang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8879-8889

Significant progress has been made in Video Object Segmentation (VOS), the video object tracking task in its finest level. While the VOS task can be naturally decoupled into image semantic segmentation and video object tracking, significantly much more research effort has been made in segmentation than tracking. In this paper, we introduce "tracking-by-detection" into VOS which can coherently integrates segmentation into tracking, by proposing a new temporal aggregation network and a novel dynamic time-evolving template matching mechanism to achieve significantly improved performance. Notably, our method is entirely online and thus suitable for one-shot learning, and our end-to-end trainable model allows multiple object segmentation in one forward pass. We achieve new state-of-the-art performance on the DAVIS benchmark without complicated bells and whistles in both speed and accuracy, with a speed of 0.14 second per frame and J &F measure of 75.9% respectively.

****

Affinity Graph Supervision for Visual Recognition

Chu Wang, Babak Samari, Vladimir G. Kim, Siddhartha Chaudhuri, Kaleem Siddiqi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni

tion (CVPR), 2020, pp. 8247-8255
Affinity graphs are widely used in deep architectures, including graph convoluti onal neural networks and attention networks. Thus far, the literature has focuse d on abstracting features from such graphs, while the learning of the affinities themselves has been overlooked. Here we propose a principled method to directly supervise the learning of weights in affinity graphs, to exploit meaningful con nections between entities in the data source. Applied to a visual attention netw ork, our affinity supervision improves relationship recovery between objects, ev en without the use of manually annotated relationship labels. We further show th at affinity learning between objects boosts scene categorization performance and that the supervision of affinity can also be applied to graphs built from mini-batches, for neural network training. In an image classification task we demonst rate consistent improvement over the baseline, with diverse network architecture s and datasets.
****

Distilling Effective Supervision From Severe Label Noise
Zizhao Zhang, Han Zhang, Sercan O. Arik, Honglak Lee, Tomas Pfister; Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9294-9303
Collecting large-scale data with clean labels for supervised training of neural networks is practically challenging. Although noisy labels are usually cheap to acquire, existing methods suffer a lot from label noise. This paper targets at t he challenge of robust training at high label noise regimes. The key insight to achieve this goal is to wisely leverage a small trusted set to estimate exemplar weights and pseudo labels for noisy data in order to reuse them for supervised training. We present a holistic framework to train deep neural networks in a way that is highly invulnerable to label noise. Our method sets the new state of th e art on various types of label noise and achieves excellent performance on larg e-scale datasets with real-world label noise. For instance, on CIFAR100 with a 4 0% uniform noise ratio and only 10 trusted labeled data per class, our method ac hieves 80.2% classification accuracy, where the error rate is only 1.4% higher t han a neural network trained without label noise. Moreover, increasing the noise ratio to 80%, our method still maintains a high accuracy of 75.5%, compared to the previous best accuracy 48.2%.
****

Temporally Distributed Networks for Fast Video Semantic Segmentation
Ping Hu, Fabian Caba, Oliver Wang, Zhe Lin, Stan Sclaroff, Federico Perazzi ; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit ion (CVPR), 2020, pp. 8818-8827
We present TDNet, a temporally distributed network designed for fast and accurat e video semantic segmentation. We observe that features extracted from a certain high-level layer of a deep CNN can be approximated by composing features extrac ted from several shallower sub-networks. Leveraging the inherent temporal contin uity in videos, we distribute these sub-networks over sequential frames. Therefo re, at each time step, we only need to perform a lightweight computation to extr act a sub-features group from a single sub-network. The full features used for s egmentation are then recomposed by application of a novel attention propagation module that compensates for geometry deformation between frames. A grouped knowl edge distillation loss is also introduced to further improve the representation power at both full and sub-feature levels. Experiments on Cityscapes, CamVid, an d NYUD-v2 demonstrate that our method achieves state-of-the-art accuracy with si gnificantly faster speed and lower latency.
****

Noise Robust Generative Adversarial Networks
Takuhiro Kaneko, Tatsuya Harada; Proceedings of the IEEE/CVF Conference on Comp uter Vision and Pattern Recognition (CVPR), 2020, pp. 8404-8414
Generative adversarial networks (GANs) are neural networks that learn data distr ibutions through adversarial training. In intensive studies, recent GANs have sh own promising results for reproducing training images. However, in spite of nois e, they reproduce images with fidelity. As an alternative, we propose a novel fa

mily of GANs called noise robust GANs (NR-GANs), which can learn a clean image g enerator even when training images are noisy. In particular, NR-GANs can solve t his problem without having complete noise information (e.g., the noise distribut ion type, noise amount, or signal-noise relationship). To achieve this, we intro duce a noise generator and train it along with a clean image generator. However, without any constraints, there is no incentive to generate an image and noise s eparately. Therefore, we propose distribution and transformation constraints tha t encourage the noise generator to capture only the noise-specific components. I n particular, considering such constraints under different assumptions, we devis e two variants of NR-GANs for signal-independent noise and three variants of NR-GANs for signal-dependent noise. On three benchmark datasets, we demonstrate the effectiveness of NR-GANs in noise robust image generation. Furthermore, we show the applicability of NR-GANs in image denoising. Our code is available at https ://github.com/takuhirok/NR-GAN/.

****

DeepDeform: Learning Non-Rigid RGB-D Reconstruction With Semi-Supervised Data

Aljaz Bozic, Michael Zollhofer, Christian Theobalt, Matthias Niessner; Procee dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP R), 2020, pp. 7002-7012

Applying data-driven approaches to non-rigid 3D reconstruction has been difficul t, which we believe can be attributed to the lack of a large-scale training corp us. Unfortunately, this method fails for important cases such as highly non-rigi d deformations. We first address this problem of lack of data by introducing a n ovel semi-supervised strategy to obtain dense inter-frame correspondences from a sparse set of annotations. This way, we obtain a large dataset of 400 scenes, o ver 390,000 RGB-D frames, and 5,533 densely aligned frame pairs; in addition, we provide a test set along with several metrics for evaluation. Based on this cor pus, we introduce a data-driven non-rigid feature matching approach, which we in tegrate into an optimization-based reconstruction pipeline. Here, we propose a n ew neural network that operates on RGB-D frames, while maintaining robustness un der large non-rigid deformations and producing accurate predictions. Our approac h significantly outperforms existing non-rigid reconstruction methods that do no t use learned data terms, as well as learning-based approaches that only use sel f-supervision.

****

Learning Video Stabilization Using Optical Flow

Jiyang Yu, Ravi Ramamoorthi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8159-8167

We propose a novel neural network that infers the per-pixel warp fields for vide o stabilization from the optical flow fields of the input video. While previous learning based video stabilization methods attempt to implicitly learn frame mot ions from color videos, our method resorts to optical flow for motion analysis a nd directly learns the stabilization using the optical flow. We also propose a p ipeline that uses optical flow principal components for motion inpainting and wa rp field smoothing, making our method robust to moving objects, occlusion and op tical flow inaccuracy, which is challenging for other video stabilization method s. Our method achieves quantitatively and visually better results than the state -of-the-art optimization based and deep learning based video stabilization metho ds. Our method also gives a 3x speed improvement compared to the optimization b ased methods.

****

Breaking the Cycle - Colleagues Are All You Need

Ori Nizan, Ayellet Tal; Proceedings of the IEEE/CVF Conference on Computer Visi on and Pattern Recognition (CVPR), 2020, pp. 7860-7869

This paper proposes a novel approach to performing image-to-image translation be tween unpaired domains. Rather than relying on a cycle constraint, our method ta kes advantage of collaboration between various GANs. This results in a multi mod al method, in which multiple optional and diverse images are produced for a give n image. Our model addresses some of the shortcomings of classical GANs: (1) It is able to remove large objects, such as glasses. (2) Since it does not need to

support the cycle constraint, no irrelevant traces of the input are left on the generated image. (3) It manages to translate between domains that require large shape modifications. Our results are shown to outperform those generated by state-of-the-art methods for several challenging applications on commonly-used datasets, both qualitatively and quantitatively.
****

Circle Loss: A Unified Perspective of Pair Similarity Optimization
Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, Yichen Wei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6398-6407
This paper provides a pair similarity optimization viewpoint on deep feature learning, aiming to maximize the within-class similarity s_p and minimize the between-class similarity s_n. We find a majority of loss functions, including the triplet loss and the softmax cross-entropy loss, embed s_n and s_p into similarity pairs and seek to reduce (s_n-s_p). Such an optimization manner is inflexible, because the penalty strength on every single similarity score is restricted to be equal. Our intuition is that if a similarity score deviates far from the optimum, it should be emphasized. To this end, we simply re-weight each similarity to highlight the less-optimized similarity scores. It results in a Circle loss, which is named due to its circular decision boundary. The Circle loss has a unified formula for two elemental deep feature learning paradigms, ph i.e. , learning with class-level labels and pair-wise labels. Analytically, we show that the Circle loss offers a more flexible optimization approach towards a more definite convergence target, compared with the loss functions optimizing (s_n-s_p). Experimentally, we demonstrate the superiority of the Circle loss on a variety of deep feature learning tasks. On face recognition, person re-identification, as well as several fine-grained image retrieval datasets, the achieved performance is on par with the state of the art.
****

A Characteristic Function Approach to Deep Implicit Generative Modeling
Abdul Fatir Ansari, Jonathan Scarlett, Harold Soh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7478-7487
Implicit Generative Models (IGMs) such as GANs have emerged as effective data-driven models for generating samples, particularly images. In this paper, we formulate the problem of learning an IGM as minimizing the expected distance between characteristic functions. Specifically, we minimize the distance between characteristic functions of the real and generated data distributions under a suitably-chosen weighting distribution. This distance metric, which we term as the characteristic function distance (CFD), can be (approximately) computed with linear time-complexity in the number of samples, in contrast with the quadratic-time Maximum Mean Discrepancy (MMD). By replacing the discrepancy measure in the critic of a GAN with the CFD, we obtain a model that is simple to implement and stable to train. The proposed metric enjoys desirable theoretical properties including continuity and differentiability with respect to generator parameters, and continuity in the weak topology. We further propose a variation of the CFD in which the weighting distribution parameters are also optimized during training; this obviates the need for manual tuning, and leads to an improvement in test power relative to CFD. We demonstrate experimentally that our proposed method outperforms WGAN and MMD-GAN variants on a variety of unsupervised image generation benchmarks.
****

Bayesian Adversarial Human Motion Synthesis
Rui Zhao, Hui Su, Qiang Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6225-6234
We propose a generative probabilistic model for human motion synthesis. Our model has a hierarchy of three layers. At the bottom layer, we utilize Hidden semi-Markov Model (HSMM), which explicitly models the spatial pose, temporal transition and speed variations in motion sequences. At the middle layer, HSMM parameters are treated as random variables which are allowed to vary across data instances

in order to capture large intra- and inter-class variations. At the top layer, hyperparameters define the prior distributions of parameters, preventing the model from overfitting. By explicitly capturing the distribution of the data and parameters, our model has a more compact parameterization compared to GAN-based generative models. We formulate the data synthesis as an adversarial Bayesian inference problem, in which the distributions of generator and discriminator parameters are obtained for data synthesis. We evaluate our method through a variety of metrics, where we show advantage than other competing methods with better fidelity and diversity. We further evaluate the synthesis quality as a data augmentation method for recognition task. Finally, we demonstrate the benefit of our fully probabilistic approach in data restoration task.

****

On Positive-Unlabeled Classification in GAN

Tianyu Guo, Chang Xu, Jiajun Huang, Yunhe Wang, Boxin Shi, Chao Xu, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8385-8393

This paper defines a positive and unlabeled classification problem for standard GANs, which then leads to a novel technique to stabilize the training of the discriminator in GANs. Traditionally, real data are taken as positive while generated data are negative. This positive-negative classification criterion was kept fixed all through the learning process of the discriminator without considering the gradually improved quality of generated data, even if they could be more realistic than real data at times. In contrast, it is more reasonable to treat the generated data as unlabeled, which could be positive or negative according to their quality. The discriminator is thus a classifier for this positive and unlabeled classification problem, and we derive a new Positive-Unlabeled GAN (PUGAN). We theoretically discuss the global optimality the proposed model will achieve and the equivalent optimization goal. Empirically, we find that PUGAN can achieve comparable or even better performance than those sophisticated discriminator stabilization methods.

****

A Unified Object Motion and Affinity Model for Online Multi-Object Tracking

Junbo Yin, Wenguan Wang, Qinghao Meng, Ruigang Yang, Jianbing Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6768-6777

Current popular online multi-object tracking (MOT) solutions apply single object trackers (SOTs) to capture object motions, while often requiring an extra affinity network to associate objects, especially for the occluded ones. This brings extra computational overhead due to repetitive feature extraction for SOT and affinity computation. Meanwhile, the model size of the sophisticated affinity network is usually non-trivial. In this paper, we propose a novel MOT framework that unifies object motion and affinity model into a single network, named UMA, in order to learn a compact feature that is discriminative for both object motion and affinity measure. In particular, UMA integrates single object tracking and metric learning into a unified triplet network by means of multi-task learning. Such design brings advantages of improved computation efficiency, low memory requirement and simplified training procedure. In addition, we equip our model with a task-specific attention module, which is used to boost task-aware feature learning. The proposed UMA can be easily trained end-to-end, and is elegant - requiring only one training stage. Experimental results show that it achieves promising performance on several MOT Challenge benchmarks.

****

Image2StyleGAN++: How to Edit the Embedded Images?

Rameen Abdal, Yipeng Qin, Peter Wonka; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8296-8305

We propose Image2StyleGAN++, a flexible image editing framework with many applications. Our framework extends the recent Image2StyleGAN in three ways. First, we introduce noise optimization as a complement to the W+ latent space embedding. Our noise optimization can restore high frequency features in images and thus significantly improves the quality of reconstructed images, e.g. a big increase of

PSNR from 20 dB to 45 dB. Second, we extend the global W+ latent space embedding to enable local embeddings. Third, we combine embedding with activation tensor manipulation to perform high quality local edits along with global semantic edits on images. Such edits motivate various high quality image editing applications, e.g. image reconstruction, image inpainting, image crossover, local style transfer, image editing using scribbles, and attribute level feature transfer. Examples of the edited images are shown across the paper for visual inspection.
****

Efficient and Robust Shape Correspondence via Sparsity-Enforced Quadratic Assignment

Rui Xiang, Rongjie Lai, Hongkai Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9513-9522

In this work, we introduce a novel local pairwise descriptor and then develop a simple, effective iterative method to solve the resulting quadratic assignment through sparsity control for shape correspondence between two approximate isometric surfaces. Our pairwise descriptor is based on the stiffness and mass matrix of finite element approximation of the Laplace-Beltrami differential operator, which is local in space, sparse to represent, and extremely easy to compute while containing global information. It allows us to deal with open surfaces, partial matching, and topological perturbations robustly. To solve the resulting quadratic assignment problem efficiently, the two key ideas of our iterative algorithm are: 1) select pairs with good (approximate) correspondence as anchor points, 2) solve a regularized quadratic assignment problem only in the neighborhood of selected anchor points through sparsity control. These two ingredients can improve and increase the number of anchor points quickly while reducing the computation cost in each quadratic assignment iteration significantly. With enough high-quality anchor points, one may use various pointwise global features with reference to these anchor points to further improve the dense shape correspondence. We use various experiments to show the efficiency, quality, and versatility of our method on large data sets, patches, and point clouds (without global meshes).
****

PolarNet: An Improved Grid Representation for Online LiDAR Point Clouds Semantic Segmentation

Yang Zhang, Zixiang Zhou, Philip David, Xiangyu Yue, Zerong Xi, Boqing Gong, Hassan Foroosh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9601-9610

The requirement of fine-grained perception by autonomous driving systems has resulted in recently increased research in the online semantic segmentation of single-scan LiDAR. Emerging datasets and technological advancements have enabled researchers to benchmark this problem and improve the applicable semantic segmentation algorithms. Still, online semantic segmentation of LiDAR scans in autonomous driving applications remains challenging due to three reasons: (1) the need for near-real-time latency with limited hardware, (2) points are distributed unevenly across space, and (3) an increasing number of more fine-grained semantic classes. The combination of the aforementioned challenges motivates us to propose a new LiDAR-specific, KNN-free segmentation algorithm - PolarNet. Instead of using common spherical or bird's-eye-view projection, our polar bird's-eye-view representation balances the points per grid and thus indirectly redistributes the network's attention over the long-tailed points distribution over the radial axis in polar coordination. We find that our encoding scheme greatly increases the mIoU in three drastically different real urban LiDAR single-scan segmentation datasets while retaining ultra low latency and near real-time throughput.
****

CascadePSP: Toward Class-Agnostic and Very High-Resolution Segmentation via Global and Local Refinement

Ho Kei Cheng, Jihoon Chung, Yu-Wing Tai, Chi-Keung Tang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8890-8899

State-of-the-art semantic segmentation methods were almost exclusively trained on images within a fixed resolution range. These segmentations are inaccurate for

very high-resolution images since using bicubic upsampling of low-resolution segmentation does not adequately capture high-resolution details along object boundaries. In this paper, we propose a novel approach to address the high-resolution segmentation problem without using any high-resolution training data. The key insight is our CascadePSP network which refines and corrects local boundaries whenever possible. Although our network is trained with low-resolution segmentation data, our method is applicable to any resolution even for very high-resolution images larger than 4K. We present quantitative and qualitative studies on different datasets to show that CascadePSP can reveal pixel-accurate segmentation boundaries using our novel refinement module without any finetuning. Thus, our method can be regarded as class-agnostic. Finally, we demonstrate the application of our model to scene parsing in multi-class segmentation.

****

GHUM & GHUML: Generative 3D Human Shape and Articulated Pose Models

Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, Cristian Sminchisescu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6184-6193

We present a statistical, articulated 3D human shape modeling pipeline, within a fully trainable, modular, deep learning framework. Given high-resolution complete 3D body scans of humans, captured in various poses, together with additional closeups of their head and facial expressions, as well as hand articulation, and given initial, artist designed, gender neutral rigged quad-meshes, we train all model parameters including non-linear shape spaces based on variational auto-encoders, pose-space deformation correctives, skeleton joint center predictors, and blend skinning functions, in a single consistent learning loop. The models are simultaneously trained with all the 3d dynamic scan data (over 60,000 diverse human configurations in our new dataset) in order to capture correlations and ensure consistency of various components. Models support facial expression analysis, as well as body (with detailed hand) shape and pose estimation. We provide fully train-able generic human models of different resolutions- the moderate-resolution GHUM consisting of 10,168 vertices and the low-resolution GHUML(ite) of 3,194 vertices-, run comparisons between them, analyze the impact of different components and illustrate their reconstruction from image data. The models will be available for research.

****

Panoptic-Based Image Synthesis

Aysegul Dundar, Karan Sapra, Guilin Liu, Andrew Tao, Bryan Catanzaro; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8070-8079

Conditional image synthesis for generating photorealistic images serves various applications for content editing to content generation. Previous conditional image synthesis algorithms mostly rely on semantic maps, and often fail in complex environments where multiple instances occlude each other. We propose a panoptic aware image synthesis network to generate high fidelity and photorealistic images conditioned on panoptic maps which unify semantic and instance information. To achieve this, we efficiently use panoptic maps in convolution and upsampling layers. We show that with the proposed changes to the generator, we can improve on the previous state-of-the-art methods by generating images in complex instance interaction environments in higher fidelity and tiny objects in more details. Furthermore, our proposed method also outperforms the previous state-of-the-art methods in metrics of mean IoU (Intersection over Union), and detAP (Detection Average Precision).

****

Unity Style Transfer for Person Re-Identification

Chong Liu, Xiaojun Chang, Yi-Dong Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6887-6896

Style variation has been a major challenge for person re-identification, which aims to match the same pedestrians across different cameras. Existing works attempted to address this problem with camera-invariant descriptor subspace learning. However, there will be more image artifacts when the difference between the ima

ges taken by different cameras is larger. To solve this problem, we propose a UnityStyle adaption method, which can smooth the style disparities within the same camera and across different cameras. Specifically, we firstly create UnityGAN to learn the style changes between cameras, producing shape-stable style-unity images for each camera, which is called UnityStyle images. Meanwhile, we use UnityStyle images to eliminate style differences between different images, which makes a better match between query and gallery. Then, we apply the proposed method to Re-ID models, expecting to obtain more style-robust depth features for querying. We conduct extensive experiments on widely used benchmark datasets to evaluate the performance of the proposed framework, the results of which confirm the superiority of the proposed model.

****

Minimal Solvers for 3D Scan Alignment With Pairs of Intersecting Lines

Andre Mateus, Srikumar Ramalingam, Pedro Miraldo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7234-7244

We explore the possibility of using line intersection constraints for 3D scan registration. Typical 3D registration algorithms exploit point and plane correspondences, while line intersection constraints have not been used in the context of 3D scan registration before. Constraints from a match of pairs of intersecting lines in two 3D scans can be seen as two 3D line intersections, a plane correspondence, and a point correspondence. In this paper, we present minimal solvers that combine these different type of constraints: 1) three line intersections and one point match; 2) one line intersection and two point matches; 3) three line intersections and one plane match; 4) one line intersection and two plane matches; and 5) one line intersection, one point match, and one plane match. To use all the available solvers, we present a hybrid RANSAC loop. We propose a non-linear refinement technique using all the inliers obtained from the RANSAC. Vast experiments with simulated data and two real-data data-sets show that the use of these features and the combined solvers improve the accuracy. The code is available.

****

Distilling Knowledge From Graph Convolutional Networks

Yiding Yang, Jiayan Qiu, Mingli Song, Dacheng Tao, Xinchao Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7074-7083

Existing knowledge distillation methods focus on convolutional neural networks (CNNs), where the input samples like images lie in a grid domain, and have largely overlooked graph convolutional networks (GCN) that handle non-grid data. In this paper, we propose to our best knowledge the first dedicated approach to distilling knowledge from a pre-trained GCN model. To enable the knowledge transfer from the teacher GCN to the student, we propose a local structure preserving module that explicitly accounts for the topological semantics of the teacher. In this module, the local structure information from both the teacher and the student are extracted as distributions, and hence minimizing the distance between these distributions enables topology-aware knowledge transfer from the teacher, yielding a compact yet high-performance student model. Moreover, the proposed approach is readily extendable to dynamic graph models, where the input graphs for the teacher and the student may differ. We evaluate the proposed method on two different datasets using GCN models of different architectures, and demonstrate that our method achieves the state-of-the-art knowledge distillation performance for GCN models.

****

Learning Oracle Attention for High-Fidelity Face Completion

Tong Zhou, Changxing Ding, Shaowen Lin, Xinchao Wang, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7680-7689

High-fidelity face completion is a challenging task due to the rich and subtle facial textures involved. What makes it more complicated is the correlations between different facial components, for example, the symmetry in texture and structure between both eyes. While recent works adopted the attention mechanism to lea

rn the contextual relations among elements of the face, they have largely overlo
oked the disastrous impacts of inaccurate attention scores; in addition, they fa
il to pay sufficient attention to key facial components, the completion results
of which largely determine the authenticity of a face image. Accordingly, in thi
s paper, we design a comprehensive framework for face completion based on the U-
Net structure. Specifically, we propose a dual spatial attention module to effic
iently learn the correlations between facial textures at multiple scales; moreov
er, we provide an oracle supervision signal to the attention module to ensure th
at the obtained attention scores are reasonable. Furthermore, we take the locati
on of the facial components as prior knowledge and impose a multi-discriminator
on these regions, with which the fidelity of facial components is significantly
promoted. Extensive experiments on two high-resolution face datasets including C
elebA-HQ and Flickr-Faces-HQ demonstrate that the proposed approach outperforms
state-of-the-art methods by large margins.
****

Image Super-Resolution With Cross-Scale Non-Local Attention and Exhaustive Self-Exemplars Mining
Yiqun Mei, Yuchen Fan, Yuqian Zhou, Lichao Huang, Thomas S. Huang, Honghui Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog
nition (CVPR), 2020, pp. 5690-5699
Deep convolution-based single image super-resolution (SISR) networks embrace the
benefits of learning from large-scale external image resources for local recove
ry, yet most existing works have ignored the long-range feature-wise similaritie
s in natural images. Some recent works have successfully leveraged this intrinsi
c feature correlation by exploring non-local attention modules. However, none of
the current deep models have studied another inherent property of images: cross
-scale feature correlation. In this paper, we propose the first Cross-Scale Non-
Local (CS-NL) attention module with integration into a recurrent neural network.
By combining the new CS-NL prior with local and in-scale non-local priors in a
powerful recurrent fusion cell, we can find more cross-scale feature correlation
s within a single low-resolution (LR) image. The performance of SISR is signific
antly improved by exhaustively integrating all possible priors. Extensive experi
ments demonstrate the effectiveness of the proposed CS-NL module by setting new
state-of-the-arts on multiple SISR benchmarks.
****

On the Regularization Properties of Structured Dropout
Ambar Pal, Connor Lane, Rene Vidal, Benjamin D. Haeffele; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp.
7671-7679
Dropout and its extensions (e.g. DropBlock and DropConnect) are popular heuristi
cs for training neural networks, which have been shown to improve generalization
performance in practice. However, a theoretical understanding of their optimiza
tion and regularization properties remains elusive. Recent work shows that in th
e case of single hidden-layer linear networks, Dropout is a stochastic gradient
descent method for minimizing a regularized loss, and that the regularizer induc
es solutions that are low-rank and balanced. In this work we show that for singl
e hidden-layer linear networks, DropBlock induces spectral k-support norm regula
rization, and promotes solutions that are low-rank and have factors with equal n
orm. We also show that the global minimizer for DropBlock can be computed in clo
sed form, and that DropConnect is equivalent to Dropout. We then show that some
of these results can be extended to a general class of Dropout-strategies, and,
with some assumptions, to deep non-linear networks when Dropout is applied to th
e last layer. We verify our theoretical claims and assumptions experimentally wi
th commonly used network architectures.
****

Deep Geometric Functional Maps: Robust Feature Learning for Shape Correspondence
Nicolas Donati, Abhishek Sharma, Maks Ovsjanikov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8592-860
1
We present a novel learning-based approach for computing correspondences between

non-rigid 3D shapes. Unlike previous methods that either require extensive training data or operate on handcrafted input descriptors and thus generalize poorly across diverse datasets, our approach is both accurate and robust to changes in shape structure. Key to our method is a feature-extraction network that learns directly from raw shape geometry, combined with a novel regularized map extraction layer and loss, based on the functional map representation. We demonstrate through extensive experiments in challenging shape matching scenarios that our method can learn from less training data than existing supervised approaches and generalizes significantly better than current descriptor-based learning methods. Our source code is available at: https://github.com/LIX-shape-analysis/GeomFmaps.
****

Iteratively-Refined Interactive 3D Medical Image Segmentation With Multi-Agent Reinforcement Learning
Xuan Liao, Wenhao Li, Qisen Xu, Xiangfeng Wang, Bo Jin, Xiaoyun Zhang, Yanfeng Wang, Ya Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9394-9402
Existing automatic 3D image segmentation methods usually fail to meet the clinic use. Many studies have explored an interactive strategy to improve the image segmentation performance by iteratively incorporating user hints. However, the dynamic process for successive interactions is largely ignored. We here propose to model the dynamic process of iterative interactive image segmentation as a Markov decision process (MDP) and solve it with reinforcement learning (RL). Unfortunately, it is intractable to use single-agent RL for voxel-wise prediction due to the large exploration space. To reduce the exploration space to a tractable size, we treat each voxel as an agent with a shared voxel-level behavior strategy so that it can be solved with multi-agent reinforcement learning. An additional advantage of this multi-agent model is to capture the dependency among voxels for segmentation task. Meanwhile, to enrich the information of previous segmentations, we reserve the prediction uncertainty in the state space of MDP and derive an adjustment action space leading to a more precise and finer segmentation. In addition, to improve the efficiency of exploration, we design a relative cross-entropy gain-based reward to update the policy in a constrained direction. Experimental results on various medical datasets have shown that our method significantly outperforms existing state-of-the-art methods, with the advantage of less interactions and a faster convergence.
****

Editing in Style: Uncovering the Local Semantics of GANs
Edo Collins, Raja Bala, Bob Price, Sabine Susstrunk; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5771-5780
While the quality of GAN image synthesis has improved tremendously in recent years, our ability to control and condition the output is still limited. Focusing on StyleGAN, we introduce a simple and effective method for making local, semantically-aware edits to a target output image. This is accomplished by borrowing elements from a source image, also a GAN output, via a novel manipulation of style vectors. Our method requires neither supervision from an external model, nor involves complex spatial morphing operations. Instead, it relies on the emergent disentanglement of semantic objects that is learned by StyleGAN during its training. Semantic editing is demonstrated on GANs producing human faces, indoor scenes, cats, and cars. We measure the locality and photorealism of the edits produced by our method, and find that it accomplishes both.
****

A Graduated Filter Method for Large Scale Robust Estimation
Huu Le, Christopher Zach; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5559-5568
Due to the highly non-convex nature of large-scale robust parameter estimation, avoiding poor local minima is challenging in real-world applications where input data is contaminated by a large or unknown fraction of outliers. In this paper, we introduce a novel solver for robust estimation that possesses a strong ability to escape poor local minima. Our algorithm is built upon the class of traditi

onal graduated optimization techniques, which are considered state-of-the-art lo cal methods to solve problems having many poor minima. The novelty of our work l ies in the introduction of an adaptive kernel (or residual) scaling scheme, whic h allows us to achieve faster convergence rates. Like other existing methods tha t aim to return good local minima for robust estimation tasks, our method relaxe s the original robust problem, but adapts a filter framework from non-linear con strained optimization to automatically choose the level of relaxation. Experimen tal results on real large-scale datasets such as bundle adjustment instances dem onstrate that our proposed method achieves competitive results.
****

Discovering Synchronized Subsets of Sequences: A Large Scale Solution
Evangelos Sariyanidi, Casey J. Zampella, Keith G. Bartley, John D. Herrington , Theodore D. Satterthwaite, Robert T. Schultz, Birkan Tunc; Proceedings of t he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9493-9502
Finding the largest subset of sequences (i.e., time series) that are correlated above a certain threshold, within large datasets, is of significant interest for computer vision and pattern recognition problems across domains, including beha vior analysis, computational biology, neuroscience, and finance. Maximal clique algorithms can be used to solve this problem, but they are not scalable. We pres ent an approximate, but highly efficient and scalable, method that represents th e search space as a union of sets called epsilon-expanded clusters, one of which is theoretically guaranteed to contain the largest subset of synchronized seque nces. The method finds synchronized sets by fitting a Euclidean ball on epsilon- expanded clusters, using Jung's theorem. We validate the method on data from the three distinct domains of facial behavior analysis, finance, and neuroscience, where we respectively discover the synchrony among pixels of face videos, stock market item prices, and dynamic brain connectivity data. Experiments show that o ur method produces results comparable to, but up to 300 times faster than, maxim al clique algorithms, with speed gains increasing exponentially with the number of input sequences.
****

DeepCap: Monocular Human Performance Capture Using Weak Supervision
Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, Christian T heobalt; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern R ecognition (CVPR), 2020, pp. 5052-5063
Human performance capture is a highly important computer vision problem with man y applications in movie production and virtual/augmented reality. Many previous performance capture approaches either required expensive multi-view setups or di d not recover dense space-time coherent geometry with frame-to-frame corresponde nces. We propose a novel deep learning approach for monocular dense human perfor mance capture. Our method is trained in a weakly supervised manner based on mult i-view supervision completely removing the need for training data with 3D ground truth annotations. The network architecture is based on two separate networks t hat disentangle the task into a pose estimation and a non-rigid surface deformat ion step. Extensive qualitative and quantitative evaluations show that our appro ach outperforms the state of the art in terms of quality and robustness.
****

Learning Physics-Guided Face Relighting Under Directional Light
Thomas Nestmeyer, Jean-Francois Lalonde, Iain Matthews, Andreas Lehrmann; Pro ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5124-5133
Relighting is an essential step in realistically transferring objects from a cap tured image into another environment. For example, authentic telepresence in Aug mented Reality requires faces to be displayed and relit consistent with the obse rver's scene lighting. We investigate end-to-end deep learning architectures tha t both de-light and relight an image of a human face. Our model decomposes the i nput image into intrinsic components according to a diffuse physics-based image formation model. We enable non-diffuse effects including cast shadows and specul ar highlights by predicting a residual correction to the diffuse render. To trai

n and evaluate our model, we collected a portrait database of 21 subjects with v
arious expressions and poses. Each sample is captured in a controlled light stag
e setup with 32 individual light sources. Our method creates precise and believa
ble relighting results and generalizes to complex illumination conditions and ch
allenging poses, including when the subject is not looking straight at the camer
a.
****

Unsupervised Representation Learning for Gaze Estimation
Yu Yu, Jean-Marc Odobez; Proceedings of the IEEE/CVF Conference on Computer Vis
ion and Pattern Recognition (CVPR), 2020, pp. 7314-7324
Although automatic gaze estimation is very important to a large variety of appli
cation areas, it is difficult to train accurate and robust gaze models, in great
 part due to the difficulty in collecting large and diverse data (annotating 3D
gaze is expensive and existing datasets use different setups). To address this i
ssue, our main contribution in this paper is to propose an effective approach to
 learn a low dimensional gaze representation without gaze annotations, which to
the best of our best knowledge, is the first work to do so. The main idea is to
rely on a gaze redirection network and use the gaze representation difference of
 the input and target images (of the redirection network) as the redirection var
iable. A redirection loss in image domain allows the joint training of both the
redirection network and the gaze representation network. In addition, we propose
 a warping field regularization which not only provides an explicit physical mea
ning to the gaze representations but also avoids redirection distortions. Promis
ing results on few-shot gaze estimation (competitive results can be achieved wit
h as few as <= 100 calibration samples), cross-dataset gaze estimation, gaze net
work pretraining, and another task (head pose estimation) demonstrate the validi
ty of our framework.
****

Learning Better Lossless Compression Using Lossy Compression
Fabian Mentzer, Luc Van Gool, Michael Tschannen; Proceedings of the IEEE/CVF C
onference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6638-6647
We leverage the powerful lossy image compression algorithm BPG to build a lossle
ss image compression system. Specifically, the original image is first decompose
d into the lossy reconstruction obtained after compressing it with BPG and the c
orresponding residual. We then model the distribution of the residual with a con
volutional neural network-based probabilistic model that is conditioned on the B
PG reconstruction, and combine it with entropy coding to losslessly encode the r
esidual. Finally, the image is stored using the concatenation of the bitstreams
produced by BPG and the learned residual coder. The resulting compression system
 achieves state-of-the-art performance in learned lossless full-resolution image
 compression, outperforming previous learned approaches as well as PNG, WebP, an
d JPEG2000.
****

Dynamic Hierarchical Mimicking Towards Consistent Optimization Objectives
Duo Li, Qifeng Chen; Proceedings of the IEEE/CVF Conference on Computer Vision
and Pattern Recognition (CVPR), 2020, pp. 7642-7651
While the depth of modern Convolutional Neural Networks (CNNs) surpasses that of
 the pioneering networks with a significant margin, the traditional way of appen
ding supervision only over the final classifier and progressively propagating gr
adient flow upstream remains the training mainstay. Seminal Deeply-Supervised Ne
tworks (DSN) were proposed to alleviate the difficulty of optimization arising f
rom gradient flow through a long chain. However, it is still vulnerable to issue
s including interference to the hierarchical representation generation process a
nd inconsistent optimization objectives, as illustrated theoretically and empiri
cally in this paper. Complementary to previous training strategies, we propose D
ynamic Hierarchical Mimicking, a generic feature learning mechanism, to advance
CNN training with enhanced generalization ability. Partially inspired by DSN, we
 fork delicately designed side branches from the intermediate layers of a given
neural network. Each branch can emerge from certain locations of the main branch
 dynamically, which not only retains representation rooted in the backbone netwo

rk but also generates more diverse representations along its own pathway. We go one step further to promote multi-level interactions among different branches through an optimization formula with probabilistic prediction matching losses, thus guaranteeing a more robust optimization process and better representation ability. Experiments on both category and instance recognition tasks demonstrate the substantial improvements of our proposed method over its corresponding counterparts using diverse state-of-the-art CNN architectures. Code and models are publicly available at https://github.com/d-li14/DHM.
****

UCTGAN: Diverse Image Inpainting Based on Unsupervised Cross-Space Translation
Lei Zhao, Qihang Mo, Sihuan Lin, Zhizhong Wang, Zhiwen Zuo, Haibo Chen, Wei Xing, Dongming Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5741-5750
Although existing image inpainting approaches have been able to produce visually realistic and semantically correct results, they produce only one result for each masked input. In order to produce multiple and diverse reasonable solutions, we present Unsupervised Cross-space Translation Generative Adversarial Network (called UCTGAN) which mainly consists of three network modules: conditional encoder module, manifold projection module and generation module. The manifold projection module and the generation module are combined to learn one-to-one image mapping between two spaces in an unsupervised way by projecting instance image space and conditional completion image space into common low-dimensional manifold space, which can greatly improve the diversity of the repaired samples. For understanding of global information, we also introduce a new cross semantic attention layer that exploits the long-range dependencies between the known parts and the completed parts, which can improve realism and appearance consistency of repaired samples. Extensive experiments on various datasets such as CelebA-HQ, Places2, Paris Street View and ImageNet clearly demonstrate that our method not only generates diverse inpainting solutions from the same image to be repaired, but also has high image quality.
****

Reciprocal Learning Networks for Human Trajectory Prediction
Hao Sun, Zhiqun Zhao, Zhihai He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7416-7425
We observe that the human trajectory is not only forward predictable, but also backward predictable. Both forward and backward trajectories follow the same social norms and obey the same physical constraints with the only difference in their time directions. Based on this unique property, we develop a new approach, called reciprocal learning, for human trajectory prediction. Two networks, forward and backward prediction networks, are tightly coupled, satisfying the reciprocal constraint, which allows them to be jointly learned. Based on this constraint, we borrow the concept of adversarial attacks of deep neural networks, which iteratively modifies the input of the network to match the given or forced network output, and develop a new method for network prediction, called reciprocal attack for matched prediction. It further improves the prediction accuracy. Our experimental results on benchmark datasets demonstrate that our new method outperforms the state-of-the-art methods for human trajectory prediction.
****

Towards Universal Representation Learning for Deep Face Recognition
Yichun Shi, Xiang Yu, Kihyuk Sohn, Manmohan Chandraker, Anil K. Jain; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6817-6826
Recognizing wild faces is extremely hard as they appear with all kinds of variations. Traditional methods either train with specifically annotated variation data from target domains, or by introducing unlabeled target variation data to adapt from the training data. Instead, we propose a universal representation learning framework that can deal with larger variation unseen in the given training data without leveraging target domain knowledge. We firstly synthesize training data alongside some semantically meaningful variations, such as low resolution, occlusion and head pose. However, directly feeding the augmented data for training

will not converge well as the newly introduced samples are mostly hard examples. We propose to split the feature embedding into multiple sub-embeddings, and associate different confidence values for each sub-embedding to smooth the training procedure. The sub-embeddings are further decorrelated by regularizing variation classification loss and variation adversarial loss on different partitions of them. Experiments show that our method achieves top performance on general face recognition datasets such as LFW and MegaFace, while significantly better on extreme benchmarks such as TinyFace and IJB-S.
****

Minimal Solutions to Relative Pose Estimation From Two Views Sharing a Common Direction With Unknown Focal Length
Yaqing Ding, Jian Yang, Jean Ponce, Hui Kong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7045-7053
We propose minimal solutions to relative pose estimation problem from two views sharing a common direction with unknown focal length. This is relevant for cameras equipped with an IMU (inertial measurement unit), e.g., smart phones, tablets. Similar to the 6-point algorithm for two cameras with unknown but equal focal lengths and 7-point algorithm for two cameras with different and unknown focal lengths, we derive new 4- and 5-point algorithms for these two cases, respectively. The proposed algorithms can cope with coplanar points, which is a degenerate configuration for these 6- and 7-point counterparts. We present a detailed analysis and comparisons with the state of the art. Experimental results on both synthetic data and real images from a smart phone demonstrate the usefulness of the proposed algorithms.
****

Deep Fair Clustering for Visual Learning
Peizhao Li, Han Zhao, Hongfu Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9070-9079
Fair clustering aims to hide sensitive attributes during data partition by balancing the distribution of protected subgroups in each cluster. Existing work attempts to address this problem by reducing it to a classical balanced clustering with a constraint on the proportion of protected subgroups of the input space. However, the input space may limit the clustering performance, and so far only low-dimensional datasets have been considered. In light of these limitations, in this paper, we propose Deep Fair Clustering (DFC) to learn fair and clustering-favorable representations for clustering simultaneously. Our approach could effectively filter out sensitive attributes from representations, and also lead to representations that are amenable for the following cluster analysis. Theoretically, we show that our fairness constraint in DFC will not incur much loss in terms of several clustering metrics. Empirically, we provide extensive experimental demonstrations on four visual datasets to corroborate the superior performance of the proposed approach over existing fair clustering and deep clustering methods on both cluster validity and fairness criterion.
****

Rotation Consistent Margin Loss for Efficient Low-Bit Face Recognition
Yudong Wu, Yichao Wu, Ruihao Gong, Yuanhao Lv, Ken Chen, Ding Liang, Xiaolin Hu, Xianglong Liu, Junjie Yan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6866-6876
In this paper, we consider the low-bit quantization problem of face recognition (FR) under the open-set protocol. Different from well explored low-bit quantization on closed-set image classification task, the open-set task is more sensitive to quantization errors (QEs). We redefine the QEs in angular space and disentangle it into class error and individual error. These two parts correspond to inter-class separability and intra-class compactness, respectively. Instead of eliminating the entire QEs, we propose the rotation consistent margin (RCM) loss to minimize the individual error, which is more essential to feature discriminative power. Extensive experiments on popular benchmark datasets such as MegaFace Challenge, Youtube Faces (YTF), Labeled Face in the Wild (LFW) and IJB-C show the superiority of proposed loss in low-bit FR quantization tasks.
****

Super-BPD: Super Boundary-to-Pixel Direction for Fast Image Segmentation
Jianqiang Wan, Yang Liu, Donglai Wei, Xiang Bai, Yongchao Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9253-9262
Image segmentation is a fundamental vision task and still remains a crucial step for many applications. In this paper, we propose a fast image segmentation method based on a novel super boundary-to-pixel direction (super-BPD) and a customized segmentation algorithm with super-BPD. Precisely, we define BPD on each pixel as a two-dimensional unit vector pointing from its nearest boundary to the pixel. In the BPD, nearby pixels from different regions have opposite directions departing from each other, and nearby pixels in the same region have directions pointing to the other or each other (i.e., around medial points). We make use of such property to partition image into super-BPDs, which are novel informative superpixels with robust direction similarity for fast grouping into segmentation regions. Extensive experimental results on BSDS500 and Pascal Context demonstrate the accuracy and efficiency of the proposed super-BPD in segmenting images. Specifically, we achieve comparable or superior performance with MCG while running at 25fps vs 0.07fps. Super-BPD also exhibits a noteworthy transferability to unseen scenes.
****
TransMoMo: Invariance-Driven Unsupervised Video Motion Retargeting
Zhuoqian Yang, Wentao Zhu, Wayne Wu, Chen Qian, Qiang Zhou, Bolei Zhou, Chen Change Loy; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5306-5315
We present a lightweight video motion retargeting approach TransMoMo that is capable of transferring motion of a person in a source video realistically to another video of a target person. Without using any paired data for supervision, the proposed method can be trained in an unsupervised manner by exploiting invariance properties of three orthogonal factors of variation including motion, structure, and view-angle. Specifically, with loss functions carefully derived based on invariance, we train an auto-encoder to disentangle the latent representations of such factors given the source and target video clips. This allows us to selectively transfer motion extracted from the source video seamlessly to the target video in spite of structural and view-angle disparities between the source and the target. The relaxed assumption of paired data allows our method to be trained on a vast amount of videos needless of manual annotation of source-target pairing, leading to improved robustness against large structural variations and extreme motion in videos. We demonstrate the effectiveness of our method over the state-of-the-art methods. Code, model and data are publicly available on our project page (https://yzhq97.github.io/transmomo).
****
D3Feat: Joint Learning of Dense Detection and Description of 3D Local Features
Xuyang Bai, Zixin Luo, Lei Zhou, Hongbo Fu, Long Quan, Chiew-Lan Tai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6359-6367
A successful point cloud registration often lies on robust establishment of sparse matches through discriminative 3D local features. Despite the fast evolution of learning-based 3D feature descriptors, little attention has been drawn to the learning of 3D feature detectors, even less for a joint learning of the two tasks. In this paper, we leverage a 3D fully convolutional network for 3D point clouds, and propose a novel and practical learning mechanism that densely predicts both a detection score and a description feature for each 3D point. In particular, we propose a keypoint selection strategy that overcomes the inherent density variations of 3D point clouds, and further propose a self-supervised detector loss guided by the on-the-fly feature matching results during training. Finally, our method achieves state-of-the-art results in both indoor and outdoor scenarios, evaluated on 3DMatch and KITTI datasets, and shows its strong generalization ability on the ETH dataset. Towards practical use, we show that by adopting a reliable feature detector, sampling a smaller number of features is sufficient to achieve accurate and fast point cloud alignment.

****
Cross-Batch Memory for Embedding Learning
Xun Wang, Haozhi Zhang, Weilin Huang, Matthew R. Scott; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6388-6397
Mining informative negative instances are of central importance to deep metric learning (DML). However, the hard-mining ability of existing DML methods is intrinsically limited by mini-batch training, where only a mini-batch of instances are accessible at each iteration. In this paper, we identify a "slow drift" phenomena by observing that the embedding features drift exceptionally slow even as the model parameters are updating throughout the training process. It suggests that the features of instances computed at preceding iterations can considerably approximate to their features extracted by current model. We propose a cross-batch memory (XBM) mechanism that memorizes the embeddings of past iterations, allowing the model to collect sufficient hard negative pairs across multiple mini-batches - even over the whole dataset. Our XBM can be directly integrated into general pair-based DML framework.We demonstrate that, without bells and whistles, XBM augmented DML can boost the performance considerably on image retrieval. In particular, with XBM, a simple contrastive loss can have large R@1 improvements of 12%-22.5% on three large-scale datasets, easily surpassing the most sophisticated state-of-the-art methods [38, 27, 2], by a large margin. Our XBM is conceptually simple, easy to implement - using several lines of codes, and is memory efficient - with a negligible 0.2 GB extra GPU memory.
****
Hierarchical Pyramid Diverse Attention Networks for Face Recognition
Qiangchang Wang, Tianyi Wu, He Zheng, Guodong Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8326-8335
Deep learning has achieved a great success in face recognition (FR), however, few existing models take hierarchical multi-scale local features into consideration. In this work, we propose a hierarchical pyramid diverse attention (HPDA) network. First, it is observed that local patches would play important roles in FR when the global face appearance changes dramatically. Some recent works apply attention modules to locate local patches automatically without relying on face landmarks. Unfortunately, without considering diversity, some learned attentions tend to have redundant responses around some similar local patches, while neglecting other potential discriminative facial parts. Meanwhile, local patches may appear at different scales due to pose variations or large expression changes. To alleviate these challenges, we propose a pyramid diverse attention (PDA) to learn multi-scale diverse local representations automatically and adaptively. More specifically, a pyramid attention is developed to capture multi-scale features. Meanwhile, a diverse learning is developed to encourage models to focus on different local patches and generate diverse local features. Second, almost all existing models focus on extracting features from the last convolutional layer, lacking of local details or small-scale face parts in lower layers. Instead of simple concatenation or addition, we propose to use a hierarchical bilinear pooling (HBP) to fuse information from multiple layers effectively. Thus, the HPDA is developed by integrating the PDA into the HBP. Experimental results on several datasets show the effectiveness of the HPDA, compared to the state-of-the-art methods.
****
ARShadowGAN: Shadow Generative Adversarial Network for Augmented Reality in Single Light Scenes
Daquan Liu, Chengjiang Long, Hongpan Zhang, Hanning Yu, Xinzhi Dong, Chunxia Xiao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8139-8148
Generating virtual object shadows consistent with the real-world environment shading effects is important but challenging in computer vision and augmented reality applications. To address this problem, we propose an end-to-end Generative Adversarial Network for shadow generation named ARShadowGAN for augmented reality in single light scenes. Our ARShadowGAN makes full use of attention mechanism an

d is able to directly model the mapping relation between the virtual object shad ow and the real-world environment without any explicit estimation of the illumin ation and 3D geometric information. In addition, we collect an image set which p rovides rich clues for shadow generation and construct a dataset for training an d evaluating our proposed ARShadowGAN. The extensive experimental results show t hat our proposed ARShadowGAN is capable of directly generating plausible virtual object shadows in single light scenes. Our source code is available at https:// github.com/ldq9526/ARShadowGAN.
****

Going Deeper With Lean Point Networks
Eric-Tuan Le, Iasonas Kokkinos, Niloy J. Mitra; Proceedings of the IEEE/CVF Co nference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9503-9512
In this work we introduce Lean Point Networks (LPNs) to train deeper and more ac curate point processing networks by relying on three novel point processing bloc ks that improve memory consumption, inference time, and accuracy: a convolution- type block for point sets that blends neighborhood information in a memory-effic ient manner; a crosslink block that efficiently shares information across low- a nd high-resolution processing branches; and a multi-resolution point cloud proce ssing block for faster diffusion of information. By combining these blocks, we d esign wider and deeper point-based architectures. We report systematic accuracy and memory consumption improvements on multiple publicly available segmentation tasks by using our generic modules as drop-in replacements for the blocks of mul tiple architectures (PointNet++, DGCNN, SpiderNet, PointCNN).
****

Semantic Image Manipulation Using Scene Graphs
Helisa Dhamo, Azade Farshad, Iro Laina, Nassir Navab, Gregory D. Hager, Fed erico Tombari, Christian Rupprecht; Proceedings of the IEEE/CVF Conference on C omputer Vision and Pattern Recognition (CVPR), 2020, pp. 5213-5222
Image manipulation can be considered a special case of image generation where th e image to be produced is a modification of an existing image. Image generation and manipulation have been, for the most part, tasks that operate on raw pixels. However, the remarkable progress in learning rich image and object representati ons has opened the way for tasks such as text-to-image or layout-to-image genera tion that are mainly driven by semantics. In our work, we address the novel prob lem of image manipulation from scene graphs, in which a user can edit images by merely applying changes in the nodes or edges of a semantic graph that is genera ted from the image. Our goal is to encode image information in a given constella tion and from there on generate new constellations, such as replacing objects or even changing relationships between objects, while respecting the semantics and style from the original image. We introduce a spatio-semantic scene graph netwo rk that does not require direct supervision for constellation changes or image e dits. This makes it possible to train the system from existing real-world datase ts with no additional annotation effort.
****

Neural Voxel Renderer: Learning an Accurate and Controllable Rendering Tool
Konstantinos Rematas, Vittorio Ferrari; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5417-5427
We present a neural rendering framework that maps a voxelized scene into a high quality image. Highly-textured objects and scene element interactions are realis tically rendered by our method, despite having a rough representation as an inpu t. Moreover, our approach allows controllable rendering: geometric and appearanc e modifications in the input are accurately propagated to the output. The user c an move, rotate and scale an object, change its appearance and texture or modify the position of the light and all these edits are represented in the final rend ering. We demonstrate the effectiveness of our approach by rendering scenes with varying appearance, from single color per object to complex, high-frequency tex tures. We show that our rerendering network can generate very detailed images th at represent precisely the appearance of the input scene. Our experiments illust rate that our approach achieves more accurate image synthesis results compared t o alternatives and can also handle low voxel grid resolutions. Finally, we show

how our neural rendering framework can capture and faithfully render objects fro
m real images and from a diverse set of classes.
****

How to Train Your Deep Multi-Object Tracker
Yihong Xu, Aljosa Osep, Yutong Ban, Radu Horaud, Laura Leal-Taixe, Xavier A
lameda-Pineda; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat
tern Recognition (CVPR), 2020, pp. 6787-6796
The recent trend in vision-based multi-object tracking (MOT) is heading towards
leveraging the representational power of deep learning to jointly learn to detec
t and track objects. However, existing methods train only certain sub-modules us
ing loss functions that often do not correlate with established tracking evaluat
ion measures such as Multi-Object Tracking Accuracy (MOTA) and Precision (MOTP).
As these measures are not differentiable, the choice of appropriate loss functi
ons for end-to-end training of multi-object tracking methods is still an open re
search problem. In this paper, we bridge this gap by proposing a differentiable
proxy of MOTA and MOTP, which we combine in a loss function suitable for end-to-
end training of deep multi-object trackers. As a key ingredient, we propose a De
ep Hungarian Net (DHN) module that approximates the Hungarian matching algorithm
. DHN allows estimating the correspondence between object tracks and ground trut
h objects to compute differentiable proxies of MOTA and MOTP, which are in turn
used to optimize deep trackers directly. We experimentally demonstrate that the
proposed differentiable framework improves the performance of existing multi-obj
ect trackers, and we establish a new state of the art on the MOTChallenge benchm
ark. Our code is publicly available from https://github.com/yihongXU/deepMOT.
****

Cascaded Deep Monocular 3D Human Pose Estimation With Evolutionary Training Data
Shichao Li, Lei Ke, Kevin Pratama, Yu-Wing Tai, Chi-Keung Tang, Kwang-Ting
Cheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec
ognition (CVPR), 2020, pp. 6173-6183
End-to-end deep representation learning has achieved remarkable accuracy for mon
ocular 3D human pose estimation, yet these models may fail for unseen poses with
limited and fixed training data. This paper proposes a novel data augmentation
method that: (1) is scalable for synthesizing massive amount of training data (o
ver 8 million valid 3D human poses with corresponding 2D projections) for traini
ng 2D-to-3D networks, (2) can effectively reduce dataset bias. Our method evolve
s a limited dataset to synthesize unseen 3D human skeletons based on a hierarchi
cal human representation and heuristics inspired by prior knowledge. Extensive e
xperiments show that our approach not only achieves state-of-the-art accuracy on
the largest public benchmark, but also generalizes significantly better to unse
en and rare poses. Relevant files and tools are available at the project website
.
****

An End-to-End Edge Aggregation Network for Moving Object Segmentation
Prashant W. Patil, Kuldeep M. Biradar, Akshay Dudhane, Subrahmanyam Murala; P
roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
(CVPR), 2020, pp. 8149-8158
Moving object segmentation in videos (MOS) is a highly demanding task for securi
ty-based applications like automated outdoor video surveillance. Most of the exi
sting techniques proposed for MOS are highly depend on fine-tuning a model on th
e first frame(s) of test sequence or complicated training procedure, which leads
to limited practical serviceability of the algorithm. In this paper, the inhere
nt correlation learning-based edge extraction mechanism (EEM) and dense residual
block (DRB) are proposed for the discriminative foreground representation. The
multi-scale EEM module provides the efficient foreground edge related informatio
n (with the help of encoder) to the decoder through skip connection at subsequen
t scale. Further, the response of the optical flow encoder stream and the last E
EM module are embedded in the bridge network. The bridge network comprises of mu
lti-scale residual blocks with dense connections to learn the effective and effi
cient foreground relevant features. Finally, to generate accurate and consistent
foreground object maps, a decoder block is proposed with skip connections from

respective multi-scale EEM module feature maps and the subsequent down-sampled r
esponse of previous frame output. Specifically, the proposed network does not re
quire any pre-trained models or fine-tuning of the parameters with the initial f
rame(s) of the test video. The performance of the proposed network is evaluated
with different configurations like disjoint, cross-data, and global training-tes
ting techniques. The ablation study is conducted to analyse each model of the pr
oposed network. To demonstrate the effectiveness of the proposed framework, a co
mprehensive analysis on four benchmark video datasets is conducted. Experimental
 results show that the proposed approach outperforms the state-of-the-art method
s for MOS.
****

Overcoming Multi-Model Forgetting in One-Shot NAS With Diversity Maximization
Miao Zhang, Huiqi Li, Shirui Pan, Xiaojun Chang, Steven Su; Proceedings of t
he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020,
pp. 7809-7818

One-Shot Neural Architecture Search (NAS) significantly improves the computation
al efficiency through weight sharing. However, this approach also introduces mul
ti-model forgetting during the supernet training (architecture search phase), wh
ere the performance of previous architectures degrade when sequentially training
 new architectures with partially-shared weights. To overcome such catastrophic
forgetting, the state-of-the-art method assumes that the shared weights are opti
mal when jointly optimizing a posterior probability. However, this strict assump
tion is not necessarily held for One-Shot NAS in practice. In this paper, we for
mulate the supernet training in the One-Shot NAS as a constrained optimization p
roblem of continual learning that the learning of current architecture should no
t degrade the performance of previous architectures during the supernet training
. We propose a Novelty Search based Architecture Selection (NSAS) loss function
and demonstrate that the posterior probability could be calculated without the s
trict assumption when maximizing the diversity of the selected constraints. A gr
eedy novelty search method is devised to find the most representative subset to
regularize the supernet training. We apply our proposed approach to two One-Shot
 NAS baselines, random sampling NAS (RandomNAS) and gradient-based sampling NAS
(GDAS). Extensive experiments demonstrate that our method enhances the predictiv
e ability of the supernet in One-Shot NAS and achieves remarkable performance on
 CIFAR-10, CIFAR-100, and PTB with efficiency.
****

Fine-Grained Image-to-Image Transformation Towards Visual Recognition
Wei Xiong, Yutong He, Yixuan Zhang, Wenhan Luo, Lin Ma, Jiebo Luo; Proceedi
ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
, 2020, pp. 5840-5849

Existing image-to-image transformation approaches primarily focus on synthesizin
g visually pleasing data. Generating images with correct identity labels is chal
lenging yet much less explored. It is even more challenging to deal with image t
ransformation tasks with large deformation in poses, viewpoints, or scales while
 preserving the identity, such as face rotation and object viewpoint morphing. I
n this paper, we aim at transforming an image with a fine-grained category to sy
nthesize new images that preserve the identity of the input image, which can the
reby benefit the subsequent fine-grained image recognition and few-shot learning
 tasks. The generated images, transformed with large geometric deformation, do n
ot necessarily need to be of high visual quality but are required to maintain as
 much identity information as possible. To this end, we adopt a model based on g
enerative adversarial networks to disentangle the identity related and unrelated
 factors of an image. In order to preserve the fine-grained contextual details o
f the input image during the deformable transformation, a constrained nonalignme
nt connection method is proposed to construct learnable highways between interme
diate convolution blocks in the generator. Moreover, an adaptive identity modula
tion mechanism is proposed to transfer the identity information into the output
image effectively. Extensive experiments on the CompCars and Multi-PIE datasets
demonstrate that our model preserves the identity of the generated images much b
etter than the state-of-the-art image-to-image transformation models, and as a r

esult significantly boosts the visual recognition performance in fine-grained few-shot learning.
****

Self-Supervised Learning of Pretext-Invariant Representations
Ishan Misra, Laurens van der Maaten; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6707-6717
The goal of self-supervised learning from images is to construct image representations that are semantically meaningful via pretext tasks that do not require semantic annotations. Many pretext tasks lead to representations that are covariant with image transformations. We argue that, instead, semantic representations ought to be invariant under such transformations. Specifically, we develop Pretext-Invariant Representation Learning (PIRL, pronounced as `pearl') that learns invariant representations based on pretext tasks. We use PIRL with a commonly used pretext task that involves solving jigsaw puzzles. We find that PIRL substantially improves the semantic quality of the learned image representations. Our approach sets a new state-of-the-art in self-supervised learning from images on several popular benchmarks for self-supervised learning. Despite being unsupervised, PIRL outperforms supervised pre-training in learning image representations for object detection. Altogether, our results demonstrate the potential of self-supervised representations with good invariance properties.
****

HyperSTAR: Task-Aware Hyperparameters for Deep Networks
Gaurav Mittal, Chang Liu, Nikolaos Karianakis, Victor Fragoso, Mei Chen, Yun Fu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8736-8745
While deep neural networks excel in solving visual recognition tasks, they require significant effort to find hyperparameters that make them work optimally. Hyperparameter Optimization (HPO) approaches have automated the process of finding good hyperparameters but they do not adapt to a given task (task-agnostic), making them computationally inefficient. To reduce HPO time, we present HyperSTAR (System for Task Aware Hyperparameter Recommendation), a task-aware method to warm-start HPO for deep neural networks. HyperSTAR ranks and recommends hyperparameters by predicting their performance conditioned on a joint dataset-hyperparameter space. It learns a dataset (task) representation along with the performance predictor directly from raw images in an end-to-end fashion. The recommendations, when integrated with an existing HPO method, make it task-aware and significantly reduce the time to achieve optimal performance. We conduct extensive experiments on 10 publicly available large-scale image classification datasets over two different network architectures, validating that HyperSTAR evaluates 50% less configurations to achieve the best performance compared to existing methods. We further demonstrate that HyperSTAR makes Hyperband (HB) task-aware, achieving the optimal accuracy in just 25% of the budget required by both vanilla HB and Bayesian Optimized HB (BOHB).
****

Deblurring Using Analysis-Synthesis Networks Pair
Adam Kaufman, Raanan Fattal; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5811-5820
Blind image deblurring remains a challenging problem for modern artificial neural networks. Unlike other image restoration problems, deblurring networks fail behind the performance of existing deblurring algorithms in case of uniform and 3D blur models. This follows from the diverse and profound effect that the unknown blur-kernel has on the deblurring operator. We propose a new architecture which breaks the deblurring network into an analysis network which estimates the blur, and a synthesis network that uses this kernel to deblur the image. Unlike existing deblurring networks, this design allows us to explicitly incorporate the blur-kernel in the network's training. In addition, we introduce new cross-correlation layers that allow better blur estimations, as well as unique components that allow the estimate blur to control the action of the synthesis deblurring action. Evaluating the new approach over established benchmark datasets shows its ability to achieve state-of-the-art deblurring accuracy on various tests, as well

as offer a major speedup in runtime.
****
A Novel Recurrent Encoder-Decoder Structure for Large-Scale Multi-View Stereo Reconstruction From an Open Aerial Dataset
Jin Liu, Shunping Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6050-6059
A great deal of research has demonstrated recently that multi-view stereo (MVS) matching can be solved with deep learning methods. However, these efforts were focused on close-range objects and only a very few of the deep learning-based methods were specifically designed for large-scale 3D urban reconstruction due to the lack of multi-view aerial image benchmarks. In this paper, we present a synthetic aerial dataset, called the WHU dataset, we created for MVS tasks, which, to our knowledge, is the first large-scale multi-view aerial dataset. It was generated from a highly accurate 3D digital surface model produced from thousands of real aerial images with precise camera parameters. We also introduce in this paper a novel network, called RED-Net, for wide-range depth inference, which we developed from a recurrent encoder-decoder structure to regularize cost maps across depths and a 2D fully convolutional network as framework. RED-Net's low memory requirements and high performance make it suitable for large-scale and highly accurate 3D Earth surface reconstruction. Our experiments confirmed that not only did our method exceed the current state-of-the-art MVS methods by more than 50% mean absolute error (MAE) with less memory and computational cost, but its efficiency as well. It outperformed one of the best commercial software programs based on conventional methods, improving their efficiency 16 times over. Moreover, we proved that our RED-Net model pre-trained on the synthetic WHU dataset can be efficiently transferred to very different multi-view aerial image datasets without any fine-tuning. Dataset and code are available at http://gpcv.whu.edu.cn/data.
****
Deep Polarization Cues for Transparent Object Segmentation
Agastya Kalra, Vage Taamazyan, Supreeth Krishna Rao, Kartik Venkataraman, Ramesh Raskar, Achuta Kadambi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8602-8611
Segmentation of transparent objects is a hard, open problem in computer vision. Transparent objects lack texture of their own, adopting instead the texture of scene background. This paper reframes the problem of transparent object segmentation into the realm of light polarization, i.e., the rotation of light waves. We use a polarization camera to capture multi-modal imagery and couple this with a unique deep learning backbone for processing polarization input data. Our method achieves instance segmentation on cluttered, transparent objects in various scene and background conditions, demonstrating an improvement over traditional image-based approaches. As an application we use this for robotic bin picking of transparent objects.
****
GAN Compression: Efficient Architectures for Interactive Conditional GANs
Muyang Li, Ji Lin, Yaoyao Ding, Zhijian Liu, Jun-Yan Zhu, Song Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5284-5294
Conditional Generative Adversarial Networks (cGANs) have enabled controllable image synthesis for many computer vision and graphics applications. However, recent cGANs are 1-2 orders of magnitude more computationally-intensive than modern recognition CNNs. For example, GauGAN consumes 281G MACs per image, compared to 0.44G MACs for MobileNet-v3, making it difficult for interactive deployment. In this work, we propose a general-purpose compression framework for reducing the inference time and model size of the generator in cGANs. Directly applying existing CNNs compression methods yields poor performance due to the difficulty of GAN training and the differences in generator architectures. We address these challenges in two ways. First, to stabilize the GAN training, we transfer knowledge of multiple intermediate representations of the original model to its compressed model, and unify unpaired and paired learning. Second, instead of reusing existin

g CNN designs, our method automatically finds efficient architectures via neural architecture search (NAS). To accelerate the search process, we decouple the model training and architecture search via weight sharing. Experiments demonstrate the effectiveness of our method across different supervision settings (paired and unpaired), model architectures, and learning methods (e.g., pix2pix, GauGAN, CycleGAN). Without losing image quality, we reduce the computation of CycleGAN by more than 20x and GauGAN by 9x, paving the way for interactive image synthesis. The code and demo are publicly available.
****

Joint Training of Variational Auto-Encoder and Latent Energy-Based Model
Tian Han, Erik Nijkamp, Linqi Zhou, Bo Pang, Song-Chun Zhu, Ying Nian Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7978-7987
This paper proposes a joint training method to learn both the variational auto-encoder (VAE) and the latent energy-based model (EBM). The joint training of VAE and latent EBM are based on an objective function that consists of three Kullback-Leibler divergences between three joint distributions on the latent vector and the image, and the objective function is of an elegant symmetric and anti-symmetric form of divergence triangle that seamlessly integrates variational and adversarial learning. In this joint training scheme, the latent EBM serves as a critic of the generator model, while the generator model and the inference model in VAE serve as the approximate synthesis sampler and inference sampler of the latent EBM. Our experiments show that the joint training greatly improves the synthesis quality of the VAE. It also enables learning of an energy function that is capable of detecting out of sample examples for anomaly detection.
****

Data-Efficient Semi-Supervised Learning by Reliable Edge Mining
Peibin Chen, Tao Ma, Xu Qin, Weidi Xu, Shuchang Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9192-9201
Learning powerful discriminative features is a challenging task in Semi-Supervised Learning, as the estimation of the feature space is more likely to be wrong with scarcer labeled data. Previous methods utilize a relation graph with edges representing 'similarity' or 'dissimilarity' between nodes. Similar nodes are forced to output consistent features, while dissimilar nodes are forced to be inconsistent. However, since unlabeled data may be wrongly labeled, the judgment of edges may be unreliable. Besides, the nodes connected by edges may already be well fitted, thus contributing little to the model training. We propose Reliable Edge Mining (REM), which forms a reliable graph by only selecting reliable and useful edges. Guided by the graph, the feature extractor is able to learn discriminative features in a data-efficient way, and consequently boosts the accuracy of the learned classifier. Visual analyses show that the features learned are more discriminative and better reveals the underlying structure of the data. REM can be combined with perturbation-based methods like Pi-model, TempEns and Mean Teacher to further improve accuracy. Experiments prove that our method is data-efficient on simple tasks like SVHN and CIFAR-10, and achieves state-of-the-art results on the challenging CIFAR-100.
****

Stylization-Based Architecture for Fast Deep Exemplar Colorization
Zhongyou Xu, Tingting Wang, Faming Fang, Yun Sheng, Guixu Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9363-9372
Exemplar-based colorization aims to add colors to a grayscale image guided by a content related reference im- age. Existing methods are either sensitive to the selection of reference images (content, position) or extremely time and resource consuming, which limits their practical applica- tion. To tackle these problems, we propose a deep exemplar colorization architecture inspired by the characteristics of stylization in feature extracting and blending. Our coarse- to-fine architecture consists of two parts: a fast transfer sub-net and a robust colorization sub-net. The transfer sub- net obtains a coarse chrominance map via matching

basic feature statistics of the input pairs in a progressive way. The colorizat ion sub-net refines the map to generate the final re- sults. The proposed end-to -end network can jointly learn faithful colorization with a related reference an d plausible color prediction with unrelated reference. Extensive exper- imental validation demonstrates that our approach outper- forms the state-of-the-art met hods in less time whether in exemplar-based colorization or image stylization ta sks.
****
PSGAN: Pose and Expression Robust Spatial-Aware GAN for Customizable Makeup Tran sfer
Wentao Jiang, Si Liu, Chen Gao, Jie Cao, Ran He, Jiashi Feng, Shuicheng Ya n; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni tion (CVPR), 2020, pp. 5194-5202
In this paper, we address the makeup transfer task, which aims to transfer the m akeup from a reference image to a source image. Existing methods have achieved p romising progress in constrained scenarios, but transferring between images with large pose and expression differences is still challenging. Besides, they canno t realize customizable transfer that allows a controllable shade of makeup or sp ecifies the part to transfer, which limits their applications. To address these issues, we propose Pose and expression robust Spatial-aware GAN (PSGAN). It firs t utilizes Makeup Distill Network to disentangle the makeup of the reference ima ge as two spatial-aware makeup matrices. Then, Attentive Makeup Morphing module is introduced to specify how the makeup of a pixel in the source image is morphe d from the reference image. With the makeup matrices and the source image, Makeu p Apply Network is used to perform makeup transfer. Our PSGAN not only achieves state-of-the-art results even when large pose and expression differences exist b ut also is able to perform partial and shade-controllable makeup transfer. Both the code and a newly collected dataset containing facial images with various pos es and expressions will be available at https://github.com/wtjiang98/PSGAN.
****
Spatial Pyramid Based Graph Reasoning for Semantic Segmentation
Xia Li, Yibo Yang, Qijie Zhao, Tiancheng Shen, Zhouchen Lin, Hong Liu; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2020, pp. 8950-8959
The convolution operation suffers from a limited receptive filed, while global m odeling is fundamental to dense prediction tasks, such as semantic segmentation. In this paper, we apply graph convolution into the semantic segmentation task a nd propose an improved Laplacian. The graph reasoning is directly performed in t he original feature space organized as a spatial pyramid. Different from existin g methods, our Laplacian is data-dependent and we introduce an attention diagona l matrix to learn a better distance metric. It gets rid of projecting and re-pro jecting processes, which makes our proposed method a light-weight module that ca n be easily plugged into current computer vision architectures. More importantly , performing graph reasoning directly in the feature space retains spatial relat ionships and makes spatial pyramid possible to explore multiple long-range conte xtual patterns from different scales. Experiments on Cityscapes, COCO Stuff, PAS CAL Context and PASCAL VOC demonstrate the effectiveness of our proposed methods on semantic segmentation. We achieve comparable performance with advantages in computational and memory overhead.
****
GAMIN: Generative Adversarial Multiple Imputation Network for Highly Missing Dat a
Seongwook Yoon, Sanghoon Sull; Proceedings of the IEEE/CVF Conference on Comput er Vision and Pattern Recognition (CVPR), 2020, pp. 8456-8464
We propose a novel imputation method for highly missing data. Though most existi ng imputation methods focus on moderate missing rate, imputation for high missin g rate over 80% is still important but challenging. As we expect that multiple i mputation is indispensable for high missing rate, we propose a generative advers arial multiple imputation network (GAMIN) based on generative adversarial networ k (GAN) for multiple imputation. Compared with similar imputation methods adopti

ng GAN, our method has three novel contributions: 1)We propose a novel imputatio n architecture which generates candidates of imputation. 2)We present a confiden ce prediction method to perform reliable multiple imputation. 3)We realize them with GAMIN and train it using novel loss functions based on the confidence. We s ynthesized highly missing datasets using MNIST and CelebA to perform various exp eriments. The results show that our method outperforms baseline methods at high missing rate from 80% to 95%.
****
When to Use Convolutional Neural Networks for Inverse Problems
Nathaniel Chodosh, Simon Lucey; Proceedings of the IEEE/CVF Conference on Compu ter Vision and Pattern Recognition (CVPR), 2020, pp. 8226-8235
Reconstruction tasks in computer vision aim fundamentally to recover an undeterm ined signal from a set of noisy measurements. Examples include super-resolution, image denoising, and non-rigid structure from motion, all of which have seen re cent advancements through deep learning. However, earlier work made extensive us e of sparse signal reconstruction frameworks (e.g. convolutional sparse coding). While this work was ultimately surpassed by deep learning, it rested on a much more developed theoretical framework. Recent work by Papyan et. al. provides a b ridge between the two approaches by showing how a convolutional neural network (CNN) can be viewed as an approximate solution to a convolutional sparse coding (CSC) problem. In this work we argue that for some types of inverse problems the CNN approximation breaks down leading to poor performance. We argue that for the se types of problems the CSC approach should be used instead and validate this a rgument with empirical evidence. Specifically we identify JPEG artifact reductio n and non-rigid trajectory reconstruction as challenging inverse problems for CN Ns and demonstrate state of the art performance on them using a CSC method. Furt hermore, we offer some practical improvements to this model and its application, and also show how insights from the CSC model can be used to make CNNs effectiv e in tasks where their naive application fails.
****
Dynamic Face Video Segmentation via Reinforcement Learning
Yujiang Wang, Mingzhi Dong, Jie Shen, Yang Wu, Shiyang Cheng, Maja Pantic; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio n (CVPR), 2020, pp. 6959-6969
For real-time semantic video segmentation, most recent works utilised a dynamic framework with a key scheduler to make online key/non-key decisions. Some works used a fixed key scheduling policy, while others proposed adaptive key schedulin g methods based on heuristic strategies, both of which may lead to suboptimal gl obal performance. To overcome this limitation, we model the online key decision process in dynamic video segmentation as a deep reinforcement learning problem a nd learn an efficient and effective scheduling policy from expert information ab out decision history and from the process of maximising global return. Moreover, we study the application of dynamic video segmentation on face videos, a field that has not been investigated before. By evaluating on the 300VW dataset, we sh ow that the performance of our reinforcement key scheduler outperforms that of v arious baselines in terms of both effective key selections and running speed. Fu rther results on the Cityscapes dataset demonstrate that our proposed method can also generalise to other scenarios. To the best of our knowledge, this is the f irst work to use reinforcement learning for online key-frame decision in dynamic video segmentation, and also the first work on its application on face videos.
****
ManiGAN: Text-Guided Image Manipulation
Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, Philip H.S. Torr; Proceedings of t he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7880-7889
The goal of our paper is to semantically edit parts of an image matching a given text that describes desired attributes (e.g., texture, colour, and background), while preserving other contents that are irrelevant to the text. To achieve thi s, we propose a novel generative adversarial network (ManiGAN), which contains t wo key components: text-image affine combination module (ACM) and detail correct

ion module (DCM). The ACM selects image regions relevant to the given text and t
hen correlates the regions with corresponding semantic words for effective manip
ulation. Meanwhile, it encodes original image features to help reconstruct text-
irrelevant contents. The DCM rectifies mismatched attributes and completes missi
ng contents of the synthetic image. Finally, we suggest a new metric for evaluat
ing image manipulation results, in terms of both the generation of new attribute
s and the reconstruction of text-irrelevant contents. Extensive experiments on t
he CUB and COCO datasets demonstrate the superior performance of the proposed me
thod.
****

The GAN That Warped: Semantic Attribute Editing With Unpaired Data
Garoe Dorta, Sara Vicente, Neill D. F. Campbell, Ivor J. A. Simpson; Proceedi
ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
, 2020, pp. 5356-5365
Deep neural networks have recently been used to edit images with great success,
in particular for faces. However, they are often limited to only being able to w
ork at a restricted range of resolutions. Many methods are so flexible that face
 edits can often result in an unwanted loss of identity. This work proposes to l
earn how to perform semantic image edits through the application of smooth warp
fields. Previous approaches that attempted to use warping for semantic edits req
uired paired data, i.e. example images of the same subject with different semant
ic attributes. In contrast, we employ recent advances in Generative Adversarial
Networks that allow our model to be trained with unpaired data. We demonstrate f
ace editing at very high resolutions (4k images) with a single forward pass of a
 deep network at a lower resolution. We also show that our edits are substantial
ly better at preserving the subject's identity. The robustness of our approach i
s demonstrated by showing plausible image editing results on the Cub200 birds da
taset. To our knowledge this has not been previously accomplished, due the chall
enging nature of the dataset.
****

Detecting Attended Visual Targets in Video
Eunji Chong, Yongxin Wang, Nataniel Ruiz, James M. Rehg; Proceedings of the I
EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp.
5396-5406
We address the problem of detecting attention targets in video. Our goal is to i
dentify where each person in each frame of a video is looking, and correctly han
dle the case where the gaze target is out-of-frame. Our novel architecture model
s the dynamic interaction between the scene and head features and infers time-va
rying attention targets. We introduce a new annotated dataset, VideoAttentionTar
get, containing complex and dynamic patterns of real-world gaze behavior. Our ex
periments show that our model can effectively infer dynamic attention in videos.
 In addition, we apply our predicted attention maps to two social gaze behavior
recognition tasks, and show that the resulting classifiers significantly outperf
orm existing methods. We achieve state-of-the-art performance on three datasets:
 GazeFollow (static images), VideoAttentionTarget (videos), and VideoCoAtt (vide
os), and obtain the first results for automatically classifying clinically-relev
ant gaze behavior without wearable cameras or eye trackers.
****

Total Deep Variation for Linear Inverse Problems
Erich Kobler, Alexander Effland, Karl Kunisch, Thomas Pock; Proceedings of th
e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, p
p. 7549-7558
Diverse inverse problems in imaging can be cast as variational problems composed
 of a task-specific data fidelity term and a regularization term. In this paper,
 we propose a novel learnable general-purpose regularizer exploiting recent arch
itectural design patterns from deep learning. We cast the learning problem as a
discrete sampled optimal control problem, for which we derive the adjoint state
equations and an optimality condition. By exploiting the variational structure o
f our approach, we perform a sensitivity analysis with respect to the learned pa
rameters obtained from different training datasets. Moreover, we carry out a non

linear eigenfunction analysis, which reveals interesting properties of the learned regularizer. We show state-of-the-art performance for classical image restoration and medical image reconstruction problems.
****

Learning Multi-Object Tracking and Segmentation From Automatic Annotations
Lorenzo Porzi, Markus Hofinger, Idoia Ruiz, Joan Serrat, Samuel Rota Bulo, Peter Kontschieder; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6846-6855
In this work we contribute a novel pipeline to automatically generate training data, and to improve over state-of-the-art multi-object tracking and segmentation (MOTS) methods. Our proposed track mining algorithm turns raw street-level videos into high-fidelity MOTS training data, is scalable and overcomes the need of expensive and time-consuming manual annotation approaches. We leverage state-of-the-art instance segmentation results in combination with optical flow predictions, also trained on automatically harvested training data. Our second major contribution is MOTSNet - a deep learning, tracking-by-detection architecture for MOTS - deploying a novel mask-pooling layer for improved object association over time. Training MOTSNet with our automatically extracted data leads to significantly improved sMOTSA scores on the novel KITTI MOTS dataset (+1.9%/+7.5% on cars/pedestrians), and MOTSNet improves by +4.1% over previously best methods on the MOTSChallenge dataset. Our most impressive finding is that we can improve over previous best-performing works, even in complete absence of manually annotated MOTS training data.
****

GeoDA: A Geometric Framework for Black-Box Adversarial Attacks
Ali Rahmati, Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard, Huaiyu Dai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8446-8455
Adversarial examples are known as carefully perturbed images fooling image classifiers. We propose a geometric framework to generate adversarial examples in one of the most challenging black-box settings where the adversary can only generate a small number of queries, each of them returning the top-1 label of the classifier. Our framework is based on the observation that the decision boundary of deep networks usually has a small mean curvature in the vicinity of data samples. We propose an effective iterative algorithm to generate query-efficient black-box perturbations with small p norms which is confirmed via experimental evaluations on state-of-the-art natural image classifiers. Moreover, for p=2, we theoretically show that our algorithm actually converges to the minimal perturbation when the curvature of the decision boundary is bounded. We also obtain the optimal distribution of the queries over the iterations of the algorithm. Finally, experimental results confirm that our principled black-box attack algorithm performs better than state-of-the-art algorithms as it generates smaller perturbations with a reduced number of queries.
****

Semantically Multi-Modal Image Synthesis
Zhen Zhu, Zhiliang Xu, Ansheng You, Xiang Bai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5467-5476
In this paper, we focus on semantically multi-modal image synthesis (SMIS) task, namely, generating multi-modal images at the semantic level. Previous work seeks to use multiple class-specific generators, constraining its usage in datasets with a small number of classes. We instead propose a novel Group Decreasing Network (GroupDNet) that leverages group convolutions in the generator and progressively decreases the group numbers of the convolutions in the decoder. Consequently, GroupDNet is armed with much more controllability on translating semantic labels to natural images and has plausible high-quality yields for datasets with many classes. Experiments on several challenging datasets demonstrate the superiority of GroupDNet on performing the SMIS task. We also show that GroupDNet is capable of performing a wide range of interesting synthesis applications. Codes and models are available at: https://github.com/Seanseattle/SMIS.
****

Copy and Paste GAN: Face Hallucination From Shaded Thumbnails
Yang Zhang, Ivor W. Tsang, Yawei Luo, Chang-Hui Hu, Xiaobo Lu, Xin Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7355-7364
Existing face hallucination methods based on convolutional neural networks (CNN) have achieved impressive performance on low-resolution (LR) faces in a normal illumination condition. However, their performance degrades dramatically when LR faces are captured in low or non-uniform illumination conditions. This paper proposes a Copy and Paste Generative Adversarial Network (CPGAN) to recover authentic high-resolution (HR) face images while compensating for low and non-uniform illumination. To this end, we develop two key components in our CPGAN: internal and external Copy and Paste nets (CPnets). Specifically, our internal CPnet exploits facial information residing in the input image to enhance facial details; while our external CPnet leverages an external HR face for illumination compensation. A new illumination compensation loss is thus developed to capture illumination from the external guided face image effectively. Furthermore, our method offsets illumination and upsamples facial details alternatively in a coarse-to-fine fashion, thus alleviating the correspondence ambiguity between LR inputs and external HR inputs. Extensive experiments demonstrate that our method manifests authentic HR face images in a uniform illumination condition and outperforms state-of-the-art methods qualitatively and quantitatively.
****
Leveraging 2D Data to Learn Textured 3D Mesh Generation
Paul Henderson, Vagia Tsiminaki, Christoph H. Lampert; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7498-7507
Numerous methods have been proposed for probabilistic generative modelling of 3D objects. However, none of these is able to produce textured objects, which renders them of limited use for practical tasks. In this work, we present the first generative model of textured 3D meshes. Training such a model would traditionally require a large dataset of textured meshes, but unfortunately, existing datasets of meshes lack detailed textures. We instead propose a new training methodology that allows learning from collections of 2D images without any 3D information. To do so, we train our model to explain a distribution of images by modelling each image as a 3D foreground object placed in front of a 2D background. Thus, it learns to generate meshes that when rendered, produce images similar to those in its training set. A well-known problem when generating meshes with deep networks is the emergence of self-intersections, which are problematic for many use-cases. As a second contribution we therefore introduce a new generation process for 3D meshes that guarantees no self-intersections arise, based on the physical intuition that faces should push one another out of the way as they move. We conduct extensive experiments on our approach, reporting quantitative and qualitative results on both synthetic data and natural images. These show our method successfully learns to generate plausible and diverse textured 3D samples for five challenging object classes.
****
Bidirectional Graph Reasoning Network for Panoptic Segmentation
Yangxin Wu, Gengwei Zhang, Yiming Gao, Xiajun Deng, Ke Gong, Xiaodan Liang, Liang Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9080-9089
Recent researches on panoptic segmentation resort to a single end-to-end network to combine the tasks of instance segmentation and semantic segmentation. However, prior models only unified the two related tasks at the architectural level via a multi-branch scheme or revealed the underlying correlation between them by unidirectional feature fusion, which disregards the explicit semantic and co-occurrence relations among objects and background. Inspired by the fact that context information is critical to recognize and localize the objects, and inclusive object details are significant to parse the background scene, we thus investigate on explicitly modeling the correlations between object and background to achieve a holistic understanding of an image in the panoptic segmentation task. We intr

oduce a Bidirectional Graph Reasoning Network (BGRNet), which incorporates graph structure into the conventional panoptic segmentation network to mine the intra-modular and inter-modular relations within and between foreground things and background stuff classes. In particular, BGRNet first constructs image-specific graphs in both instance and semantic segmentation branches that enable flexible reasoning at the proposal level and class level, respectively. To establish the correlations between separate branches and fully leverage the complementary relations between things and stuff, we propose a Bidirectional Graph Connection Module to diffuse information across branches in a learnable fashion. Experimental results demonstrate the superiority of our BGRNet that achieves the new state-of-the-art performance on challenging COCO and ADE20K panoptic segmentation benchmarks.

****

High-Order Information Matters: Learning Relation and Topology for Occluded Person Re-Identification

Guan'an Wang, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, Jian Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6449-6458

Occluded person re-identification (ReID) aims to match occluded person images to holistic ones across dis-joint cameras. In this paper, we propose a novel framework by learning high-order relation and topology information for discriminative features and robust alignment. At first, we use a CNN backbone to learn feature maps and key-points estimation model to extract semantic local features. Even so, occluded images still suffer from occlusion and outliers. Then, we view the extracted local features of an image as nodes of a graph and propose an adaptive direction graph convolutional (ADGC) layer to pass relation information between nodes. The proposed ADGC layer can automatically suppress the message passing of meaningless features by dynamically learning direction and degree of linkage. When aligning two groups of local features, we view it as a graph matching problem and propose a cross-graph embedded-alignment (CGEA) layer to joint learn and embed topology information to local features, and straightly predict similarity score. The proposed CGEA layer can both take full use of alignment learned by graph matching and replace sensitive one-to-one alignment with a robust soft one. Finally, extensive experiments on occluded, partial, and holistic ReID tasks show the effectiveness of our proposed method. Specifically, our framework significantly outperforms state-of-the-art by 6.5% mAP scores on Occluded-Duke dataset.

****

Self-Supervised Monocular Scene Flow Estimation

Junhwa Hur, Stefan Roth; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7396-7405

Scene flow estimation has been receiving increasing attention for 3D environment perception. Monocular scene flow estimation - obtaining 3D structure and 3D motion from two temporally consecutive images - is a highly ill-posed problem, and practical solutions are lacking to date. We propose a novel monocular scene flow method that yields competitive accuracy and real-time performance. By taking an inverse problem view, we design a single convolutional neural network (CNN) that successfully estimates depth and 3D motion simultaneously from a classical optical flow cost volume. We adopt self-supervised learning with 3D loss functions and occlusion reasoning to leverage unlabeled data. We validate our design choices, including the proxy loss and augmentation setup. Our model achieves state-of-the-art accuracy among unsupervised/self-supervised learning approaches to monocular scene flow, and yields competitive results for the optical flow and monocular depth estimation sub-tasks. Semi-supervised fine-tuning further improves the accuracy and yields promising results in real-time.

****

End-to-End Model-Free Reinforcement Learning for Urban Driving Using Implicit Affordances

Marin Toromanoff, Emilie Wirbel, Fabien Moutarde; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7153-7162

Reinforcement Learning (RL) aims at learning an optimal behavior policy from its own experiments and not rule-based control methods. However, there is no RL algorithm yet capable of handling a task as difficult as urban driving. We present a novel technique, coined implicit affordances, to effectively leverage RL for urban driving thus including lane keeping, pedestrians and vehicles avoidance, and traffic light detection. To our knowledge we are the first to present a successful RL agent handling such a complex task especially regarding the traffic light detection. Furthermore, we have demonstrated the effectiveness of our method by winning the Camera Only track of the CARLA challenge.
****

Model Adaptation: Unsupervised Domain Adaptation Without Source Data
Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, Si Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9641-9650
In this paper, we investigate a challenging unsupervised domain adaptation setting --- unsupervised model adaptation. We aim to explore how to rely only on unlabeled target data to improve performance of an existing source prediction model on the target domain, since labeled source data may not be available in some real-world scenarios due to data privacy issues. For this purpose, we propose a new framework, which is referred to as collaborative class conditional generative adversarial net to bypass the dependence on the source data. Specifically, the prediction model is to be improved through generated target-style data, which provides more accurate guidance for the generator. As a result, the generator and the prediction model can collaborate with each other without source data. Furthermore, due to the lack of supervision from source data, we propose a weight constraint that encourages similarity to the source model. A clustering-based regularization is also introduced to produce more discriminative features in the target domain. Compared to conventional domain adaptation methods, our model achieves superior performance on multiple adaptation tasks with only unlabeled target data, which verifies its effectiveness in this challenging setting.
****

VecRoad: Point-Based Iterative Graph Exploration for Road Graphs Extraction
Yong-Qiang Tan, Shang-Hua Gao, Xuan-Yi Li, Ming-Ming Cheng, Bo Ren; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8910-8918
Extracting road graphs from aerial images automatically is more efficient and costs less than from field acquisition. This can be done by a post-processing step that vectorizes road segmentation predicted by CNN, but imperfect predictions will result in road graphs with low connectivity. On the other hand, iterative next move exploration could construct road graphs with better road connectivity, but often focuses on local information and does not provide precise alignment with the real road. To enhance the road connectivity while maintaining the precise alignment between the graph and real road, we propose a point-based iterative graph exploration scheme with segmentation-cues guidance and flexible steps. In our approach, we represent the location of the next move as a 'point' that unifies the representation of multiple constraints such as the direction and step size in each moving step. Information cues such as road segmentation and road junctions are jointly detected and utilized to guide the next move and achieve better alignment of roads. We demonstrate that our proposed method has a considerable improvement over state-of-the-art road graph extraction methods in terms of F-measure and road connectivity metrics on common datasets.
****

Uncertainty Based Camera Model Selection
Michal Polic, Stanislav Steidl, Cenek Albl, Zuzana Kukelova, Tomas Pajdla; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5991-6000
The quality and speed of Structure from Motion (SfM) methods depend significantly on the camera model chosen for the reconstruction. In most of the SfM pipelines, the camera model is manually chosen by the user. In this paper, we present a new automatic method for camera model selection in large scale SfM that is based

on efficient uncertainty evaluation. We first perform an extensive comparison of classical model selection based on known Information Criteria and show that they do not provide sufficiently accurate results when applied to camera model selection. Then we propose a new Accuracy-based Criterion, which evaluates an efficient approximation of the uncertainty of the estimated parameters in tested models. Using the new criterion, we design a camera model selection method and fine-tune it by machine learning. Our simulated and real experiments demonstrate a significant increase in reconstruction quality as well as a considerable speedup of the SfM process.
****

Learning a Neural 3D Texture Space From 2D Exemplars
Philipp Henzler, Niloy J. Mitra, Tobias Ritschel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8356-8364
We suggest a generative model of 2D and 3D natural textures with diversity, visual fidelity and at high computational efficiency. This is enabled by a family of methods that extend ideas from classic stochastic procedural texturing (Perlin noise) to learned, deep, non-linearities. Our model encodes all exemplars from a diverse set of textures without a need to be re-trained for each exemplar. Applications include texture interpolation, and learning 3D textures from 2D exemplars.
****

Structure-Preserving Super Resolution With Gradient Guidance
Cheng Ma, Yongming Rao, Yean Cheng, Ce Chen, Jiwen Lu, Jie Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7769-7778
Structures matter in single image super resolution (SISR). Recent studies benefiting from generative adversarial network (GAN) have promoted the development of SISR by recovering photo-realistic images. However, there are always undesired structural distortions in the recovered images. In this paper, we propose a structure-preserving super resolution method to alleviate the above issue while maintaining the merits of GAN-based methods to generate perceptual-pleasant details. Specifically, we exploit gradient maps of images to guide the recovery in two aspects. On the one hand, we restore high-resolution gradient maps by a gradient branch to provide additional structure priors for the SR process. On the other hand, we propose a gradient loss which imposes a second-order restriction on the super-resolved images. Along with the previous image-space loss functions, the gradient-space objectives help generative networks concentrate more on geometric structures. Moreover, our method is model-agnostic, which can be potentially used for off-the-shelf SR networks. Experimental results show that we achieve the best PI and LPIPS performance and meanwhile comparable PSNR and SSIM compared with state-of-the-art perceptual-driven SR methods. Visual results demonstrate our superiority in restoring structures while generating natural SR images.
****

Neural Contours: Learning to Draw Lines From 3D Shapes
Difan Liu, Mohamed Nabail, Aaron Hertzmann, Evangelos Kalogerakis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5428-5436
This paper introduces a method for learning to generate line drawings from 3D models. Our architecture incorporates a differentiable module operating on geometric features of the 3D model, and an image-based module operating on view-based shape representations. At test time, geometric and view-based reasoning are combined with the help of a neural module to create a line drawing. The model is trained on a large number of crowdsourced comparisons of line drawings. Experiments demonstrate that our method achieves significant improvements in line drawing over the state-of-the-art when evaluated on standard benchmarks, resulting in drawings that are comparable to those produced by experienced human artists.
****

An Efficient PointLSTM for Point Clouds Based Gesture Recognition
Yuecong Min, Yanxiao Zhang, Xiujuan Chai, Xilin Chen; Proceedings of the IEEE

/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5761-5770

Point clouds contain rich spatial information, which provides complementary cues for gesture recognition. In this paper, we formulate gesture recognition as an irregular sequence recognition problem and aim to capture long-term spatial correlations across point cloud sequences. A novel and effective PointLSTM is proposed to propagate information from past to future while preserving the spatial structure. The proposed PointLSTM combines state information from neighboring points in the past with current features to update the current states by a weight-shared LSTM layer. This method can be integrated into many other sequence learning approaches. In the task of gesture recognition, the proposed PointLSTM achieves state-of-the-art results on two challenging datasets (NVGesture and SHREC'17) and outperforms previous skeleton-based methods. To show its advantages in generalization, we evaluate our method on MSR Action3D dataset, and it produces competitive results with previous skeleton-based methods.

****

SCOUT: Self-Aware Discriminant Counterfactual Explanations

Pei Wang, Nuno Vasconcelos; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8981-8990

The problem of counterfactual visual explanations is considered. A new family of discriminant explanations is introduced. These produce heatmaps that attribute high scores to image regions informative of a classifier prediction but not of a counter class. They connect attributive explanations, which are based on a single heat map, to counterfactual explanations, which account for both predicted class and counter class. The latter are shown to be computable by combination of two discriminant explanations, with reversed class pairs. It is argued that self-awareness, namely the ability to produce classification confidence scores, is important for the computation of discriminant explanations, which seek to identify regions where it is easy to discriminate between prediction and counter class. This suggests the computation of discriminant explanations by the combination of three attribution maps. The resulting counterfactual explanations are optimization free and thus much faster than previous methods. To address the difficulty of their evaluation, a proxy task and set of quantitative metrics are also proposed. Experiments under this protocol show that the proposed counterfactual explanations outperform the state of the art while achieving speeds much faster, for popular networks. In a human-learning machine teaching experiment, they are also shown to improve mean student accuracy from chance level to 95%.

****

Select to Better Learn: Fast and Accurate Deep Learning Using Data Selection From Nonlinear Manifolds

Mohsen Joneidi, Saeed Vahidian, Ashkan Esmaeili, Weijia Wang, Nazanin Rahnavard, Bill Lin, Mubarak Shah; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7819-7829

Finding a small subset of data whose linear combination spans other data points, also called column subset selection problem (CSSP), is an important open problem in computer science with many applications in computer vision and deep learning. There are some studies that solve CSSP in a polynomial time complexity w.r.t. the size of the original dataset. A simple and efficient selection algorithm with a linear complexity order, referred to as spectrum pursuit (SP), is proposed that pursuits spectral components of the dataset using available sample points. The proposed non-greedy algorithm aims to iteratively find K data samples whose span is close to that of the first K spectral components of entire data. SP has no parameter to be fine tuned and this desirable property makes it problem-independent. The simplicity of SP enables us to extend the underlying linear model to more complex models such as nonlinear manifolds and graph-based models. The nonlinear extension of SP is introduced as kernel-SP (KSP). The superiority of the proposed algorithms is demonstrated in a wide range of applications.

****

Towards Photo-Realistic Virtual Try-On by Adaptively Generating-Preserving Image Content

Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, Ping Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7850-7859

Image visual try-on aims at transferring a target clothes image onto a reference person, and has become a hot topic in recent years. Prior arts usually focus on preserving the character of a clothes image (e.g. texture, logo, embroidery) when warping it to arbitrary human pose. However, it remains a big challenge to generate photo-realistic try-on images when large occlusions and human poses are presented in the reference person. To address this issue, we propose a novel visual try-on network, namely Adaptive Content Generating and Preserving Network (ACGPN). In particular, ACGPN first predicts semantic layout of the reference image that will be changed after try-on (e.g.long sleeve shirt-arm, arm-jacket), and then determines whether its image content needs to be generated or preserved according to the predicted semantic layout, leading to photo-realistic try-on and rich clothes details. ACGPN generally involves three major modules. First, a semantic layout generation module utilizes semantic segmentation of the reference image to progressively predict the desired semantic layout after try-on. Second, a clothes warping module warps clothes image according to the generated semantic layout, where a second-order difference constraint is introduced to stabilize the warping process during training.Third, an inpainting module for content fusion integrates all information (e.g. reference image, semantic layout, warped clothes) to adaptively produce each semantic part of human body. In comparison to the state-of-the-art methods, ACGPN can generate photo-realistic images with much better perceptual quality and richer fine-details.

****

Phase Consistent Ecological Domain Adaptation

Yanchao Yang, Dong Lao, Ganesh Sundaramoorthi, Stefano Soatto; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9011-9020

We introduce two criteria to regularize the optimization involved in learning a classifier in a domain where no annotated data are available, leveraging annotated data in a different domain, a problem known as unsupervised domain adaptation. We focus on the task of semantic segmentation, where annotated synthetic data are aplenty, but annotating real data is laborious. The first criterion, inspired by visual psychophysics, is that the map between the two image domains be phase-preserving. This restricts the set of possible learned maps, while enabling enough flexibility to transfer semantic information. The second criterion aims to leverage ecological statistics, or regularities in the scene which are manifest in any image of it, regardless of the characteristics of the illuminant or the imaging sensor. It is implemented using a deep neural network that scores the likelihood of each possible segmentation given a single un-annotated image. Incorporating these two priors in a standard domain adaptation framework improves performance across the board in the most common unsupervised domain adaptation benchmarks for semantic segmentation.

****

Information-Driven Direct RGB-D Odometry

Alejandro Fontan, Javier Civera, Rudolph Triebel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4929-4937

This paper presents an information-theoretic approach to point selection in direct RGB-D odometry. The aim is to select only the most informative measurements, in order to reduce the optimization problem with a minimal impact in the accuracy. It is usual practice in visual odometry/SLAM to track several hundreds of points, achieving real-time performance in high-end desktop PCs. Reducing their computational footprint will facilitate the implementation of odometry and SLAM in low-end platforms such as small robots and AR/VR glasses. Our experimental results show that our novel information-based selection criterion allows us to reduce the number of tracked points an order of magnitude (down to only 24 of them), achieving an accuracy similar to the state of the art (sometimes outperforming it) while reducing 10 times the computational demand.

****
Single-Side Domain Generalization for Face Anti-Spoofing
Yunpei Jia, Jie Zhang, Shiguang Shan, Xilin Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8484-8493
Existing domain generalization methods for face anti-spoofing endeavor to extract common differentiation features to improve the generalization. However, due to large distribution discrepancies among fake faces of different domains, it is difficult to seek a compact and generalized feature space for the fake faces. In this work, we propose an end-to-end single-side domain generalization framework (SSDG) to improve the generalization ability of face anti-spoofing. The main idea is to learn a generalized feature space, where the feature distribution of the real faces is compact while that of the fake ones is dispersed among domains but compact within each domain. Specifically, a feature generator is trained to make only the real faces from different domains undistinguishable, but not for the fake ones, thus forming a single-side adversarial learning. Moreover, an asymmetric triplet loss is designed to constrain the fake faces of different domains separated while the real ones aggregated. The above two points are integrated into a unified framework in an end-to-end training manner, resulting in a more generalized class boundary, especially good for samples from novel domains. Feature and weight normalization is incorporated to further improve the generalization ability. Extensive experiments show that our proposed approach is effective and outperforms the state-of-the-art methods on four public databases. The code is released online.
****
Optical Non-Line-of-Sight Physics-Based 3D Human Pose Estimation
Mariko Isogawa, Ye Yuan, Matthew O'Toole, Kris M. Kitani; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7013-7022
We describe a method for 3D human pose estimation from transient images (i.e., a 3D spatio-temporal histogram of photons) acquired by an optical non-line-of-sight (NLOS) imaging system. Our method can perceive 3D human pose by 'looking around corners' through the use of light indirectly reflected by the environment. We bring together a diverse set of technologies from NLOS imaging, human pose estimation and deep reinforcement learning to construct an end-to-end data processing pipeline that converts a raw stream of photon measurements into a full 3D human pose sequence estimate. Our contributions are the design of data representation process which includes (1) a learnable inverse point spread function (PSF) to convert raw transient images into a deep feature vector; (2) a neural humanoid control policy conditioned on the transient image feature and learned from interactions with a physics simulator; and (3) a data synthesis and augmentation strategy based on depth data that can be transferred to a real-world NLOS imaging system. Our preliminary experiments suggest that our method is able to generalize to real-world NLOS measurement to estimate physically-valid 3D human poses.
****
Barycenters of Natural Images Constrained Wasserstein Barycenters for Image Morphing
Dror Simon, Aviad Aberdam; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7910-7919
Image interpolation, or image morphing, refers to a visual transition between two (or more) input images. For such a transition to look visually appealing, its desirable properties are (i) to be smooth; (ii) to apply the minimal required change in the image; and (iii) to seem "real", avoiding unnatural artifacts in each image in the transition. To obtain a smooth and straightforward transition, one may adopt the well-known Wasserstein Barycenter Problem (WBP). While this approach guarantees minimal changes under the Wasserstein metric, the resulting images might seem unnatural. In this work, we propose a novel approach for image morphing that possesses all three desired properties. To this end, we define a constrained variant of the WBP that enforces the intermediate images to satisfy an image prior. We describe an algorithm that solves this problem and demonstrate it

using the sparse prior and generative adversarial networks.
****

Future Video Synthesis With Object Motion Prediction
Yue Wu, Rongrong Gao, Jaesik Park, Qifeng Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5539-5548
We present an approach to predict future video frames given a sequence of continuous video frames in the past. Instead of synthesizing images directly, our approach is designed to understand the complex scene dynamics by decoupling the background scene and moving objects. The appearance of the scene components in the future is predicted by non-rigid deformation of the background and affine transformation of moving objects. The anticipated appearances are combined to create a reasonable video in the future. With this procedure, our method exhibits much less tearing or distortion artifact compared to other approaches. Experimental results on the Cityscapes and KITTI datasets show that our model outperforms the state-of-the-art in terms of visual quality and accuracy.
****

Reference-Based Sketch Image Colorization Using Augmented-Self Reference and Dense Semantic Correspondence
Junsoo Lee, Eungyeup Kim, Yunsung Lee, Dongjun Kim, Jaehyuk Chang, Jaegul Choo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5801-5810
This paper tackles the automatic colorization task of a sketch image given an already-colored reference image. Colorizing a sketch image is in high demand in comics, animation, and other content creation applications, but it suffers from information scarcity of a sketch image. To address this, a reference image can render the colorization process in a reliable and user-driven manner. However, it is difficult to prepare for a training data set that has a sufficient amount of semantically meaningful pairs of images as well as the ground truth for a colored image reflecting a given reference (e.g., coloring a sketch of an originally blue car given a reference green car). To tackle this challenge, we propose to utilize the identical image with geometric distortion as a virtual reference, which makes it possible to secure the ground truth for a colored output image. Furthermore, it naturally provides the ground truth for dense semantic correspondence, which we utilize in our internal attention mechanism for color transfer from reference to sketch input. We demonstrate the effectiveness of our approach in various types of sketch image colorization via quantitative as well as qualitative evaluation against existing methods.
****

Collaborative Motion Prediction via Neural Motion Message Passing
Yue Hu, Siheng Chen, Ya Zhang, Xiao Gu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6319-6328
Motion prediction is essential and challenging for autonomous vehicles and social robots. One challenge of motion prediction is to model the interaction among traffic actors, which could cooperate with each other to avoid collisions or form groups. To address this challenge, we propose neural motion message passing (NMMP) to explicitly model the interaction and learn representations for directed interactions between actors. Based on the proposed NMMP, we design the motion prediction systems for two settings: the pedestrian setting and the joint pedestrian and vehicle setting. Both systems share a common pattern: we use an individual branch to model the behavior of a single actor and an interactive branch to model the interaction between actors, while with different wrappers to handle the varied input formats and characteristics. The experimental results show that both systems outperform the previous state-of-the-art methods on several existing benchmarks. Besides, we provide interpretability for interaction learning.
****

End-to-End Learnable Geometric Vision by Backpropagating PnP Optimization
Bo Chen, Alvaro Parra, Jiewei Cao, Nan Li, Tat-Jun Chin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8100-8109
Deep networks excel in learning patterns from large amounts of data. On the othe

r hand, many geometric vision tasks are specified as optimization problems. To s eamlessly combine deep learning and geometric vision, it is vital to perform lea rning and geometric optimization end-to-end. Towards this aim, we present BPnP, a novel network module that backpropagates gradients through a Perspective-n-Poi nts (PnP) solver to guide parameter updates of a neural network. Based on implic it differentiation, we show that the gradients of a "self-contained" PnP solver can be derived accurately and efficiently, as if the optimizer block were a diff erentiable function. We validate BPnP by incorporating it in a deep model that c an learn camera intrinsics, camera extrinsics (poses) and 3D structure from trai ning datasets. Further, we develop an end-to-end trainable pipeline for object p ose estimation, which achieves greater accuracy by combining feature-based heatm ap losses with 2D-3D reprojection errors. Since our approach can be extended to other optimization problems, our work helps to pave the way to perform learnable geometric vision in a principled manner. Our PyTorch implementation of BPnP is available on http://github.com/BoChenYS/BPnP.

****

Learning Texture Transformer Network for Image Super-Resolution

Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, Baining Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5791-5800

We study on image super-resolution (SR), which aims to recover realistic texture s from a low-resolution (LR) image. Recent progress has been made by taking high -resolution images as references (Ref), so that relevant textures can be transfe rred to LR images. However, existing SR approaches neglect to use attention mech anisms to transfer high-resolution (HR) textures from Ref images, which limits t hese approaches in challenging cases. In this paper, we propose a novel Texture Transformer Network for Image Super-Resolution (TTSR), in which the LR and Ref i mages are formulated as queries and keys in a transformer, respectively. TTSR co nsists of four closely-related modules optimized for image generation tasks, inc luding a learnable texture extractor by DNN, a relevance embedding module, a har d-attention module for texture transfer, and a soft-attention module for texture synthesis. Such a design encourages joint feature learning across LR and Ref im ages, in which deep feature correspondences can be discovered by attention, and thus accurate texture features can be transferred. The proposed texture transfor mer can be further stacked in a cross-scale way, which enables texture recovery from different levels (e.g., from 1x to 4x magnification). Extensive experiments show that TTSR achieves significant improvements over state-of-the-art approach es on both quantitative and qualitative evaluations.

****

Distribution-Induced Bidirectional Generative Adversarial Network for Graph Repr esentation Learning

Shuai Zheng, Zhenfeng Zhu, Xingxing Zhang, Zhizhe Liu, Jian Cheng, Yao Zhao ; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit ion (CVPR), 2020, pp. 7224-7233

Graph representation learning aims to encode all nodes of a graph into low-dimen sional vectors that will serve as input of many computer vision tasks. However, most existing algorithms ignore the existence of inherent data distribution and even noises. This may significantly increase the phenomenon of over-fitting and deteriorate the testing accuracy. In this paper, we propose a Distribution-induc ed Bidirectional Generative Adversarial Network (named DBGAN) for graph represen tation learning. Instead of the widely used Gaussian assumption, the prior distr ibution of latent representation in our DBGAN is estimated in a structure-aware way, which implicitly bridges the graph and content spaces by prototype learning . Thus discriminative and robust representations are generated for all nodes. Fu rthermore, to improve their generalization ability while preserving representati on ability, the sample-level and distribution-level consistency are well balance d via a bidirectional adversarial learning framework. An extensive group of expe riments is then carefully designed and presented, demonstrating that our DBGAN o btains remarkably more favorable trade-off between representation and robustness , and meanwhile is dimension-efficient, over currently available alternatives in

various tasks.
****

Benchmarking the Robustness of Semantic Segmentation Models
Christoph Kamann, Carsten Rother; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8828-8838
When designing a semantic segmentation module for a practical application, such as autonomous driving, it is crucial to understand the robustness of the module with respect to a wide range of image corruptions. While there are recent robustness studies for full-image classification, we are the first to present an exhaustive study for semantic segmentation, based on the state-of-the-art model DeepLabv3+. To increase the realism of our study, we utilize almost 400,000 images generated from Cityscapes, PASCAL VOC 2012, and ADE20K. Based on the benchmark study, we gain several new insights. Firstly, contrary to full-image classification, model robustness increases with model performance, in most cases. Secondly, some architecture properties affect robustness significantly, such as a Dense Prediction Cell, which was designed to maximize performance on clean data only.
****

Local Implicit Grid Representations for 3D Scenes
Chiyu "Max" Jiang, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias Niessner, Thomas Funkhouser; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6001-6010
Shape priors learned from data are commonly used to reconstruct 3D objects from partial or noisy data. Yet no such shape priors are available for indoor scenes, since typical 3D autoencoders cannot handle their scale, complexity, or diversity. In this paper, we introduce Local Implicit Grid Representations, a new 3D shape representation designed for scalability and generality. The motivating idea is that most 3D surfaces share geometric details at some scale -- i.e., at a scale smaller than an entire object and larger than a small patch. We train an autoencoder to learn an embedding of local crops of 3D shapes at that size. Then, we use the decoder as a component in a shape optimization that solves for a set of latent codes on a regular grid of overlapping crops such that an interpolation of the decoded local shapes matches a partial or noisy observation. We demonstrate the value of this proposed approach for 3D surface reconstruction from sparse point observations, showing significantly better results than alternative approaches.
****

Deformable Siamese Attention Networks for Visual Object Tracking
Yuechen Yu, Yilei Xiong, Weilin Huang, Matthew R. Scott; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6728-6737
Siamese-based trackers have achieved excellent performance on visual object tracking. However, the target template is not updated online, and the features of target template and search image are computed independently in a Siamese architecture. In this paper, we propose Deformable Siamese Attention Networks, referred to as SiamAttn, by introducing a new Siamese attention mechanism that computes deformable self-attention and cross-attention. The self-attention learns strong context information via spatial attention, and selectively emphasizes interdependent channel-wise features with channel attention. The crossattention is capable of aggregating rich contextual interdependencies between the target template and the search image, providing an implicit manner to adaptively update the target template. In addition, we design a region refinement module that computes depth-wise cross correlations between the attentional features for more accurate tracking. We conduct experiments on six benchmarks, where our method achieves new state-of-the-art results, outperforming recent strong baseline, SiamRPN++, by 0.464 to 0.537 and 0.415 to 0.470 EAO on VOT 2016 and 2018.
****

Learning Video Object Segmentation From Unlabeled Videos
Xiankai Lu, Wenguan Wang, Jianbing Shen, Yu-Wing Tai, David J. Crandall, Steven C. H. Hoi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8960-8970

We propose a new method for video object segmentation (VOS) that addresses objec t pattern learning from unlabeled videos, unlike most existing methods which rel y heavily on extensive annotated data. We introduce a unified unsupervised/weakl y supervised learning framework, called MuG, that comprehensively captures intri nsic properties of VOS at multiple granularities. Our approach can help advance understanding of visual patterns in VOS and significantly reduce annotation burd en. With a carefully-designed architecture and strong representation learning ab ility, our learned model can be applied to diverse VOS settings, including objec t-level zero-shot VOS, instance-level zero-shot VOS, and one-shot VOS. Experimen ts demonstrate promising performance in these settings, as well as the potential of MuG in leveraging unlabeled data to further improve the segmentation accurac y.

****

ScopeFlow: Dynamic Scene Scoping for Optical Flow

Aviram Bar-Haim, Lior Wolf; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7998-8007

We propose to modify the common training protocols of optical flow, leading to s izable accuracy improvements without adding to the computational complexity of t he training process. The improvement is based on observing the bias in sampling challenging data that exists in the current training protocol, and improving the sampling process. In addition, we find that both regularization and augmentatio n should decrease during the training protocol. Using an existing low parameters architecture, the method is ranked first on the MPI Sintel benchmark among all other methods, improving the best two frames method accuracy by more than 10%. T he method also surpasses all similar architecture variants by more than 12% and 19.7% on the KITTI benchmarks, achieving the lowest Average End-Point Error on K ITTI2012 among two-frame methods, without using extra datasets.

****

Context-Aware Human Motion Prediction

Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer; Procee dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP R), 2020, pp. 6992-7001

The problem of predicting human motion given a sequence of past observations is at the core of many applications in robotics and computer vision. Current state-of-the-art formulates this problem as a sequence-to-sequence task, in which a hi storical of 3D skeletons feeds a Recurrent Neural Network (RNN) that predicts fu ture movements, typically in the order of 1 to 2 seconds. However, one aspect th at has been obviated so far, is the fact that human motion is inherently driven by interactions with objects and/or other humans in the environment. In this pap er, we explore this scenario using a novel context-aware motion prediction archi tecture. We use a semantic-graph model where the nodes parameterize the human an d objects in the scene and the edges their mutual interactions. These interactio ns are iteratively learned through a graph attention layer, fed with the past ob servations, which now include both object and human body motions. Once this sema ntic graph is learned, we inject it to a standard RNN to predict future movement s of the human/s and object/s. We consider two variants of our architecture, eit her freezing the contextual interactions in the future of updating them. A thoro ugh evaluation in the Whole-Body Human Motion Database shows that in both cases, our context-aware networks clearly outperform baselines in which the context in formation is not considered.

****

MISC: Multi-Condition Injection and Spatially-Adaptive Compositing for Condition al Person Image Synthesis

Shuchen Weng, Wenbo Li, Dawei Li, Hongxia Jin, Boxin Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp . 7741-7749

In this paper, we explore synthesizing person images with multiple conditions fo r various backgrounds. To this end, we propose a framework named "MISC" for cond itional image generation and image compositing. For conditional image generation , we improve the existing condition injection mechanisms by leveraging the inter

-condition correlations. For the image compositing, we theoretically prove the w eaknesses of the cutting-edge methods, and make it more robust by removing the s patially-invariance constraint, and enabling the bounding mechanism and the spat ial adaptability. We show the effectiveness of our method on the Video Instance- level Parsing dataset, and demonstrate the robustness through controllability te sts.
****

Pathological Retinal Region Segmentation From OCT Images Using Geometric Relatio n Based Augmentation

Dwarikanath Mahapatra, Behzad Bozorgtabar, Ling Shao; Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9611 -9620

Medical image segmentation is important for computer aided diagnosis. Pixelwise manual annotations of large datasets require high expertise and is time consumin g. Conventional data augmentations have limited benefit by not fully representin g the underlying distribution of the training set, thus affecting model robustne ss when tested on images captured from different sources. Prior work leverages s ynthetic images for data augmentation ignoring the interleaved geometric relatio nship between different anatomical labels. We propose improvements over previous GAN-based medical image synthesis methods by jointly encoding the intrinsic rel ationship of geometry and shape. Latent space variable sampling results in diver se generated images from a base image and improves robustness. Augmented dataset s using our method for automatic segmentation of retinal optical coherence tomog raphy (OCT) images outperform existing methods on the public RETOUCH dataset hav ing images captured from different acquisition procedures. Ablation studies and visual analysis also demonstrate benefits of integrating geometry and diversity.
****

Filter Grafting for Deep Neural Networks

Fanxu Meng, Hao Cheng, Ke Li, Zhixin Xu, Rongrong Ji, Xing Sun, Guangming Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn ition (CVPR), 2020, pp. 6599-6607

This paper proposes a new learning paradigm called filter grafting, which aims t o improve the representation capability of Deep Neural Networks (DNNs). The moti vation is that DNNs have unimportant (invalid) filters (e.g., l1 norm close to 0). These filters limit the potential of DNNs since they are identified as having little effect on the network. While filter pruning removes these invalid filter s for efficiency consideration, filter grafting re-activates them from an accura cy boosting perspective. The activation is processed by grafting external inform ation (weights) into invalid filters. To better perform the grafting process, we develop an entropy-based criterion to measure the information of filters and an adaptive weighting strategy for balancing the grafted information among network s. After the grafting operation, the network has very few invalid filters compar ed with its untouched state, enpowering the model with more representation capac ity. We also perform extensive experiments on the classification and recognition tasks to show the superiority of our method. For example, the grafted MobileNet V2 outperforms the non-grafted MobileNetV2 by about 7 percent on CIFAR-100 datas et.
****

Intuitive, Interactive Beard and Hair Synthesis With Generative Models

Kyle Olszewski, Duygu Ceylan, Jun Xing, Jose Echevarria, Zhili Chen, Weikai Chen, Hao Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pa ttern Recognition (CVPR), 2020, pp. 7446-7456

We present an interactive approach to synthesizing realistic variations in facia l hair in images, ranging from subtle edits to existing hair to the addition of complex and challenging hair in images of clean-shaven subjects. To circumvent t he tedious and computationally expensive tasks of modeling, rendering and compos iting the 3D geometry of the target hairstyle using the traditional graphics pip eline, we employ a neural network pipeline that synthesizes realistic and detail ed images of facial hair directly in the target image in under one second. The s ynthesis is controlled by simple and sparse guide strokes from the user defining

the general structural and color properties of the target hairstyle. We qualitatively and quantitatively evaluate our chosen method compared to several alternative approaches. We show compelling interactive editing results with a prototype user interface that allows novice users to progressively refine the generated image to match their desired hairstyle, and demonstrate that our approach also allows for flexible and high-fidelity scalp hair synthesis.
****

Global Optimality for Point Set Registration Using Semidefinite Programming
Jose Pedro Iglesias, Carl Olsson, Fredrik Kahl; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8287-8295
In this paper we present a study of global optimality conditions for Point Set Registration (PSR) with missing data. PSR is the problem of aligning multiple point clouds with an unknown target point cloud. Since non-linear rotation constraints are present the problem is inherently non-convex and typically relaxed by computing the Lagrange dual, which is a Semidefinite Program (SDP). In this work we show that given a local minimizer the dual variables of the SDP can be computed in closed form. This opens up the possibility of verifying the optimally, using the SDP formulation without explicitly solving it. In addition it allows us to study under what conditions the relaxation is tight, through spectral analysis. We show that if the errors in the (unknown) optimal solution are bounded the SDP formulation will be able to recover it.
****

SQE: a Self Quality Evaluation Metric for Parameters Optimization in Multi-Object Tracking
Yanru Huang, Feiyu Zhu, Zheni Zeng, Xi Qiu, Yuan Shen, Jianan Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8306-8314
We present a novel self quality evaluation metric SQE for parameters optimization in the challenging yet critical multi-object tracking task. Current evaluation metrics all require annotated ground truth, thus will fail in the test environment and realistic circumstances prohibiting further optimization after training. By contrast, our metric reflects the internal characteristics of trajectory hypotheses and measures tracking performance without ground truth. We demonstrate that trajectories with different qualities exhibit different single or multiple peaks over feature distance distribution, inspiring us to design a simple yet effective method to assess the quality of trajectories using a two-class Gaussian mixture model. Experiments mainly on MOT16 Challenge data sets verify the effectiveness of our method in both correlating with existing metrics and enabling parameters self-optimization to achieve better performance. We believe that our conclusions and method are inspiring for future multi-object tracking in practice.
****

PointASNL: Robust Point Clouds Processing Using Nonlocal Neural Networks With Adaptive Sampling
Xu Yan, Chaoda Zheng, Zhen Li, Sheng Wang, Shuguang Cui; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5589-5598
Raw point clouds data inevitably contains outliers or noise through acquisition from 3D sensors or reconstruction algorithms. In this paper, we present a novel end-to-end network for robust point clouds processing, named PointASNL, which can deal with point clouds with noise effectively. The key component in our approach is the adaptive sampling (AS) module. It first re-weights the neighbors around the initial sampled points from farthest point sampling (FPS), and then adaptively adjusts the sampled points beyond the entire point cloud. Our AS module can not only benefit the feature learning of point clouds, but also ease the biased effect of outliers. To further capture the neighbor and long-range dependencies of the sampled point, we proposed a local-nonlocal (L-NL) module inspired by the nonlocal operation. Such L-NL module enables the learning process insensitive to noise. Extensive experiments verify the robustness and superiority of our approach in point clouds processing tasks regardless of synthesis data, indoor data, and outdoor data with or without noise. Specifically, PointASNL achieves state

-of-the-art robust performance for classification and segmentation tasks on all datasets, and significantly outperforms previous methods on real-world outdoor SemanticKITTI dataset with considerate noise.
****

Minimizing Discrete Total Curvature for Image Processing
Qiuxiang Zhong, Yutong Li, Yijie Yang, Yuping Duan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9474-9482
The curvature regularities have received growing attention with the advantage of providing strong priors in the continuity of edges in image processing applications. However, owing to the non-convex and non-smooth properties of the high-order regularizer, the numerical solution becomes challenging in real-time tasks. In this paper, we propose a novel curvature regularity, the total curvature (TC), by minimizing the normal curvatures along different directions. We estimate the normal curvatures discretely in the local neighborhood according to differential geometry theory. The resulting curvature regularity can be regarded as a re-weighted total variation (TV) minimization problem, which can be efficiently solved by the alternating direction method of multipliers (ADMM) based algorithm. By comparing with TV and Euler's elastica energy, we demonstrate the effectiveness and superiority of the total curvature regularity for various image processing applications.
****

Unsupervised Learning of Intrinsic Structural Representation Points
Nenglun Chen, Lingjie Liu, Zhiming Cui, Runnan Chen, Duygu Ceylan, Changhe Tu, Wenping Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9121-9130
Learning structures of 3D shapes is a fundamental problem in the field of computer graphics and geometry processing. We present a simple yet interpretable unsupervised method for learning a new structural representation in the form of 3D structure points. The 3D structure points produced by our method encode the shape structure intrinsically and exhibit semantic consistency across all the shape instances with similar structures. This is a challenging goal that has not fully been achieved by other methods. Specifically, our method takes a 3D point cloud as input and encodes it as a set of local features. The local features are then passed through a novel point integration module to produce a set of 3D structure points. The chamfer distance is used as reconstruction loss to ensure the structure points lie close to the input point cloud. Extensive experiments have shown that our method outperforms the state-of-the-art on the semantic shape correspondence task and achieves comparable performance with the state-of-the-art on the segmentation label transfer task. Moreover, the PCA based shape embedding built upon consistent structure points demonstrates good performance in preserving the shape structures. Code is available at https://github.com/NolenChen/3DStructurePoints
****

Deep Active Learning for Biased Datasets via Fisher Kernel Self-Supervision
Denis Gudovskiy, Alec Hodgkinson, Takuya Yamaguchi, Sotaro Tsukizawa; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9041-9049
Active learning (AL) aims to minimize labeling efforts for data-demanding deep neural networks (DNNs) by selecting the most representative data points for annotation. However, currently used methods are ill-equipped to deal with biased data. The main motivation of this paper is to consider a realistic setting for pool-based semi-supervised AL, where the unlabeled collection of train data is biased. We theoretically derive an optimal acquisition function for AL in this setting. It can be formulated as distribution shift minimization between unlabeled train data and weakly-labeled validation dataset. To implement such acquisition function, we propose a low-complexity method for feature density matching using self-supervised Fisher kernel (FK) as well as several novel pseudo-label estimators. Our FK-based method outperforms state-of-the-art methods on MNIST, SVHN, and ImageNet classification while requiring only 1/10th of processing. The conducted e

xperiments show at least 40% drop in labeling efforts for the biased class-imbal anced data compared to existing methods.
****

FGN: Fully Guided Network for Few-Shot Instance Segmentation
Zhibo Fan, Jin-Gang Yu, Zhihao Liang, Jiarong Ou, Changxin Gao, Gui-Song Xi a, Yuanqing Li; Proceedings of the IEEE/CVF Conference on Computer Vision and P attern Recognition (CVPR), 2020, pp. 9172-9181
Few-shot instance segmentation (FSIS) conjoins the few-shot learning paradigm wi th general instance segmentation, which provides a possible way of tackling inst ance segmentation in the lack of abundant labeled data for training. This paper presents a Fully Guided Network (FGN) for few-shot instance segmentation. FGN pe rceives FSIS as a guided model where a so-called support set is encoded and util ized to guide the predictions of a base instance segmentation network (i.e., Mas k R-CNN), critical to which is the guidance mechanism. In this view, FGN introdu ces different guidance mechanisms into the various key components in Mask R-CNN, including Attention-Guided RPN, Relation-Guided Detector, and Attention-Guided FCN, in order to make full use of the guidance effect from the support set and a dapt better to the inter-class generalization. Experiments on public datasets de monstrate that our proposed FGN can outperform the state-of-the-art methods.
****

DualConvMesh-Net: Joint Geodesic and Euclidean Convolutions on 3D Meshes
Jonas Schult, Francis Engelmann, Theodora Kontogianni, Bastian Leibe; Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8612-8622
We propose DualConvMesh-Nets (DCM-Net) a family of deep hierarchical convolution al networks over 3D geometric data that combines two types of convolutions. The first type, Geodesic convolutions, defines the kernel weights over mesh surfaces or graphs. That is, the convolutional kernel weights are mapped to the local su rface of a given mesh. The second type, Euclidean convolutions, is independent o f any underlying mesh structure. The convolutional kernel is applied on a neighb orhood obtained from a local affinity representation based on the Euclidean dist ance between 3D points. Intuitively, geodesic convolutions can easily separate o bjects that are spatially close but have disconnected surfaces, while Euclidean convolutions can represent interactions between nearby objects better, as they a re oblivious to object surfaces. To realize a multi-resolution architecture, we borrow well-established mesh simplification methods from the geometry processing domain and adapt them to define mesh-preserving pooling and unpooling operation s. We experimentally show that combining both types of convolutions in our archi tecture leads to significant performance gains for 3D semantic segmentation, and we report competitive results on three scene segmentation benchmarks. Models an d code will be made publicly available.
****

Noise Modeling, Synthesis and Classification for Generic Object Anti-Spoofing
Joel Stehouwer, Amin Jourabloo, Yaojie Liu, Xiaoming Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7294-7303
Using printed photograph and replaying videos of biometric modalities, such as i ris, fingerprint and face, are common attacks to fool the recognition systems fo r granting access as the genuine user. With the growing online person-to-person shopping (e.g., Ebay and Craigslist), such attacks also threaten those services, where the online photo illustration might not be captured from real items but f rom paper or digital screen. Thus, the study of anti-spoofing should be extended from modality-specific solutions to generic-object-based ones. In this work, we define and tackle the problem of Generic Object Anti-Spoofing (GOAS) for the fi rst time. One significant cue to detect these attacks is the noise patterns intr oduced by the capture sensors and spoof mediums. Different sensor/medium combina tions can result in diverse noise patterns. We propose a GAN-based architecture to synthesize and identify the noise patterns from seen and unseen medium/sensor combinations. We show that the procedure of synthesis and identification are mu tually beneficial. We further demonstrate the learned GOAS models can directly c

ontribute to modality-specific anti-spoofing without domain transfer. The code a
nd GOSet dataset are available at cvlab.cse.msu.edu/project-goas.html.
****

An Investigation Into the Stochasticity of Batch Whitening
Lei Huang, Lei Zhao, Yi Zhou, Fan Zhu, Li Liu, Ling Shao; Proceedings of th
e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, p
p. 6439-6448
Batch Normalization (BN) is extensively employed in various network architecture
s by performing standardization within mini-batches. A full understanding of the
process has been a central target in the deep learning communities. Unlike exis
ting works, which usually only analyze the standardization operation, this paper
investigates the more general Batch Whitening (BW). Our work originates from th
e observation that while various whitening transformations equivalently improve
the conditioning, they show significantly different behaviors in discriminative
scenarios and training Generative Adversarial Networks (GANs). We attribute this
phenomenon to the stochasticity that BW introduces. We quantitatively investiga
te the stochasticity of different whitening transformations and show that it cor
relates well with the optimization behaviors during training. We also investigat
e how stochasticity relates to the estimation of population statistics during in
ference. Based on our analysis, we provide a framework for designing and compari
ng BW algorithms in different scenarios. Our proposed BW algorithm improves the
residual networks by a significant margin on ImageNet classification. Besides, w
e show that the stochasticity of BW can improve the GAN's performance with, howe
ver, the sacrifice of the training stability.
****

VIBE: Video Inference for Human Body Pose and Shape Estimation
Muhammed Kocabas, Nikos Athanasiou, Michael J. Black; Proceedings of the IEEE/
CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5253
-5263
Human motion is fundamental to understanding behavior. Despite progress on singl
e-image 3D pose and shape estimation, existing video-based state-of-the-art meth
ods fail to produce accurate and natural motion sequences due to a lack of groun
d-truth 3D motion data for training. To address this problem, we propose "Video
Inference for Body Pose and Shape Estimation" (VIBE), which makes use of an exis
ting large-scale motion capture dataset (AMASS) together with unpaired, in-the-w
ild, 2D keypoint annotations. Our key novelty is an adversarial learning framewo
rk that leverages AMASS to discriminate between real human motions and those pro
duced by our temporal pose and shape regression networks. We define a novel temp
oral network architecture with a self-attention mechanism and show that adversar
ial training, at the sequence level, produces kinematically plausible motion seq
uences without in-the-wild ground-truth 3D labels. We perform extensive experime
ntation to analyze the importance of motion and demonstrate the effectiveness of
VIBE on challenging 3D pose estimation datasets, achieving state-of-the-art per
formance. Code and pretrained models are available at https://github.com/mkocaba
s/VIBE
****